DMDTEval: An Evaluation and Analysis of LLMs on Disambiguation in Multi-domain Translation

Anonymous ACL submission

Abstract

Currently, Large Language Models (LLMs) have achieved remarkable results in machine translation. However, their performance in multi-domain translation (MDT) is less satisfactory, the meanings of words can vary across different domains, highlighting the significant ambiguity inherent in MDT. Therefore, evaluating the disambiguation ability of LLMs in MDT, remains an open problem. To this end, we present an evaluation and analysis of LLMs on **d**isambiguation in **m**ulti-**d**omain **t**ranslation (**DMDTEval**), our systematic evaluation framework consisting of three critical aspects: **(1)** we construct a translation test set with multi-domain ambiguous word annotation, **(2)** we curate a diverse set of disambiguation prompt strategies, and **(3)** we design precise disambiguation metrics, and study the efficacy of various prompt strategies on multiple state-of-the-art LLMs. We conduct comprehensive experiments across 4 language pairs and 13 domains, our extensive experiments reveal a number of crucial findings that we believe will pave the way and also facilitate further research in the critical area of improving the disambiguation of LLMs. Our code and data will be released.

1 Introduction

In recent years, LLMs achieve the promising results in machine translation (MT) that demonstrate their potential in practical applications (Jiao et al., 2023b; Qian et al., 2024; Feng et al., 2025). However, LLMs perform unsatisfactorily in multi-domain translation (MDT) (Zheng et al., 2024; Hu et al., 2024b). LLMs rely on extensive pre-training data, but multi-domain parallel corpora remain exceedingly scarce. This scarcity limits their translation capabilities and prevents them from effectively acquiring cross-domain knowledge, which leads to translation ambiguities. Figure 1, Example ①, shows that directly using LLMs for translation causes word ambiguities. For example, the term

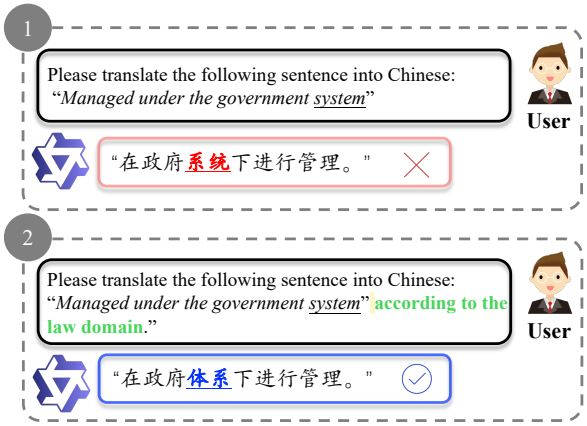

Figure 1: Two examples from the UM-Corpus English-Chinese test set. We prompt LLMs with domain label to disambiguate in Qwen-2.5-7B-Instruct. Red text represents for the ambiguity translation. Blue text represents for the correct translation (*hereinafter the same*).

"*system*," which refers to "体系" (*framework*), may be mistranslated as "系统" (*the literal translation of system*). This example illustrates that word ambiguity poses a key challenge for LLMs in MDT.

An intuitive solution is to directly prompt the LLMs to translate according to the specific domain (Hu et al., 2024b), and we find that this approach yields the correct translation. The translation of the term "*system*" in the Law domain is accurate "体系", as shown in the Figure 1, Example ②. **The critical issue is how to effectively leverage domain information in prompt strategies to enhance the performance of LLMs.**

Regarding the above critical issue, previous work mainly focuses on two key aspects: **(i) Multi-domain translation** (Jiang et al., 2020; Man et al., 2023, 2024c): these methods aim to enhance translation performance across different domains by incorporating sentence-level and word-level domain labels. Recently, some researchers have explored the performance of LLMs in MDT (Hu et al., 2024b) and investigated fine-tuning LLMs using

domain-specific parallel corpora (Hu et al., 2024b; Zheng et al., 2024). **(ii) Disambiguation evaluation for translation**: (Campolungo et al., 2022; Maheshwari et al., 2024; Martelli et al., 2025): these studies eluate the ability of models to handle and translate lexical ambiguities in general domains. The above-mentioned work provides feasible approaches for MDT under LLMs. However, three key research questions (**RQ**) remain unresolved in MDT:

- **RQ1:** ***How can we quantify the disambiguation ability of LLMs in MDT?*** Existing work (i) evaluates or fine-tunes MDT with LLMs, but does not address the role of key factors (*i.e.,* ambiguity) that influence performance variation in MDT. Therefore, constructing an ambiguity dataset and designing evaluation metrics for ambiguity are crucial.

- **RQ2:** ***Can various prompting techniques help LLMs disambiguate in MDT?*** Figure 1, Example ②, shows that the translation changes when the prompt includes domain information. This observation suggests that domain information influences the translation of LLMs. Therefore, we explore additional prompt strategies to determine how they affect the performance of LLMs in MDT.

- **RQ3**: ***What domain knowledge is essential for LLMs to achieve effective MDT?*** Previous work (ii) mainly evaluates ambiguity in general domains. In the MDT, the core research questions revolve around cross-domain word ambiguities and identifying which domain knowledge can be effectively leveraged under LLMs.

To answer and explore the aforementioned questions, we introduce an evaluation and analysis of LLMs on disambiguation in multi-domain translation for LLMs (DMDTEval) to tackle the challenges in MDT. **For RQ1**: We employ a word alignment tool to construct a multi-domain ambiguity vocabulary and manually annotate ambiguous words in the test set. Additionally, we design an evaluation metric to assess disambiguation ability in translation and compute the accuracy of ambiguous words being correctly translated. **For RQ2**: We design multiple disambiguation prompt strategiess to evaluate the translation performance of prominent LLMs across multiple domains. **For RQ3**: We conduct extensive experiments across four language pairs, with a particular focus on English-Chinese translation, providing a detailed and in-depth analysis along with key findings based on these experimental results.

To sum up, the main contributions of our work can be summarized as follows:

- We construct an ambiguous word dataset specifically tailored for MDT. This dataset enables systematic evaluation of the disambiguation capabilities of LLMs. We will open-source it to support future research on enhancing the disambiguation performance of LLMs in MDT.

- We systematically explore various disambiguation prompt strategies, including zero-shot, chain-of-thought (CoT), few-shot, and reflection prompting, to evaluate MDT quality using 5 popular open-source LLMs.

- We investigate the types of domain knowledge required by LLMs to evaluate translations across 4 language pairs and 13 domains, focusing on sentence-level and word-level domain knowledge, domain-specific examples, and domain discrimination capabilities.

2 DMDTEval: Evaluation Framework

In our work, our goal includes **(1)** constructing a n ambiguous word test set (§2.1). **(2)** evaluating the influence of domain information in LLMs' translation with different prompting (§2.2). **(3)** and designing the metrics of word ambiguity (§2.4).

2.1 Data Construction

In this section, we aim to construct a multi-domain ambiguous word vocabulary to annotate the test set. Currently, the publicly available test sets of domain-specific machine translation is scarce. We use the same dataset as in previous research (Man et al., 2024a; Hu et al., 2024b), we mainly utilize two MDT test sets for ambiguous data set[1]: UM-Corpus[2] (English-Chinese), including five domains: *Education, Law, News Science, and Spoken* (Tian

[1]Since this part of the data involves manual annotation, we primarily construct ambiguous data sets for English–Chinese and German–English, given our linguistic expertise in these language pairs. For Japanese–English and Korean–English, we utilize these data to evaluate overall translation quality.

[2]http://nlp2ct.cis.umac.mo/um-corpus/

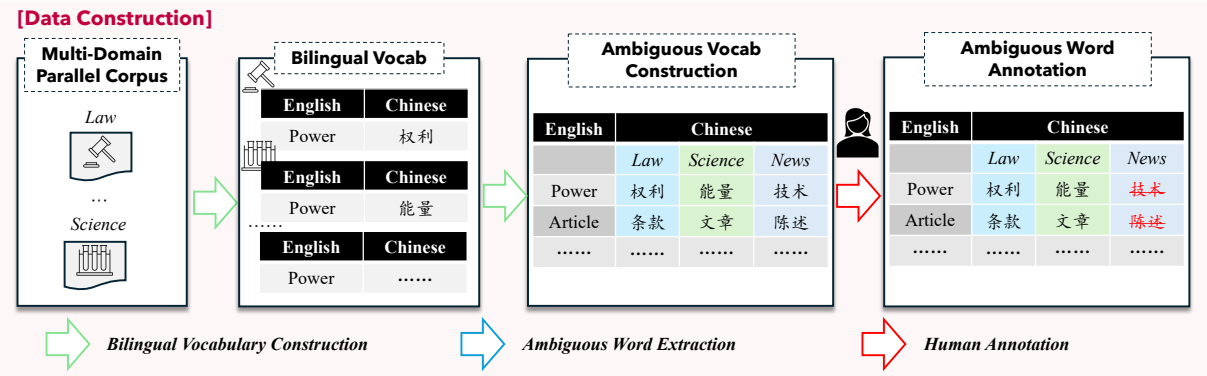

Figure 2: Ambiguous word test set construction annotation. This process consists of three steps.

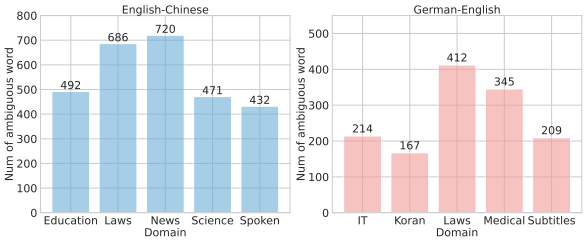

Figure 3: Statistics of ambiguous word in the test set.

et al., 2014), and OPUS[3] (German-English), including five domains: *IT, Koran, Laws, Medical, and Subtitles* (Aharoni and Goldberg, 2020). The detailed statistic of these data sets in Appendix A. We utilize the train set from these domains to obtain an ambiguity vocabulary, as shown in the Figure 2. Our annotation processing consists of three steps:

Step 1: Bilingual Vocabulary Construction. In this step, we apply *Awesome-Align*[4] (Dou and Neubig, 2021) to perform word alignment on multi-domain training corpora and extract bilingual word pairs. We then deduplicate and merge the bilingual vocabularies within each domain based on the source-language tokens. This process yields domain-specific bilingual lexicons, which include a substantial number of ambiguous words (*e.g.*, "*power*" → "权力" in Law domain, "*power*" → "能量" in Science domain).

Step 2: Ambiguous Vocabulary Construction. In this step, we construct a cross-domain ambiguous vocabulary based on the bilingual lexicons obtained in Step 1. For each domain, we initialize an empty set to store ambiguous word pairs. Then, for each bilingual pair in the domain-specific lexicon, we check whether the source word appears in other domains with different target-language translations. If such discrepancies are found, all corresponding translations are added to the ambiguous vocabulary set for that domain. This process results in a collection of domain-specific ambiguous vocabularies.

Step 3: Human Annotation. Due to inevitable errors in word alignment, we manually refine the bilingual lexicons derived from the alignment process. In this step, we annotate the sentences in each domain's test set using the ambiguous vocabulary obtained in Step 2. Specifically, we identify and label instances of one-to-many source-language words that appear in the test set. The statistics of such ambiguous words are summarized in Figure 3.

Scoring of Alignment Quality. To evaluate the quality of the word alignments in ambiguous vocabulary construction, we randomly sample aligned word pairs from each domain and ask bilingual annotators to judge their correctness. Each pair is labeled as `correct`, `partially correct`, or `incorrect`. We calculate alignment accuracy as the proportion of correct alignments. Table 7 shows the results across domains in Appendix A, highlighting the need for human annotation in Step 3.

2.2 Design of Prompt Strategies

In this section, we introduce the design of prompt templates, including both base prompt templates and disambiguation prompt templates, as shown in Figure 4.

Base Prompt Strategies. Designing an effective prompt is the key to unlocking the translation capabilities of LLMs. Specifically, we evaluate impact of different base prompt strategies, including: **(1) Zero-shot** directly asks LLM to translate a source input into the target language (Liu et al., 2018). **(2)**

[3]http://opus.nlpl.eu/
[4]https://github.com/neulab/awesome-align

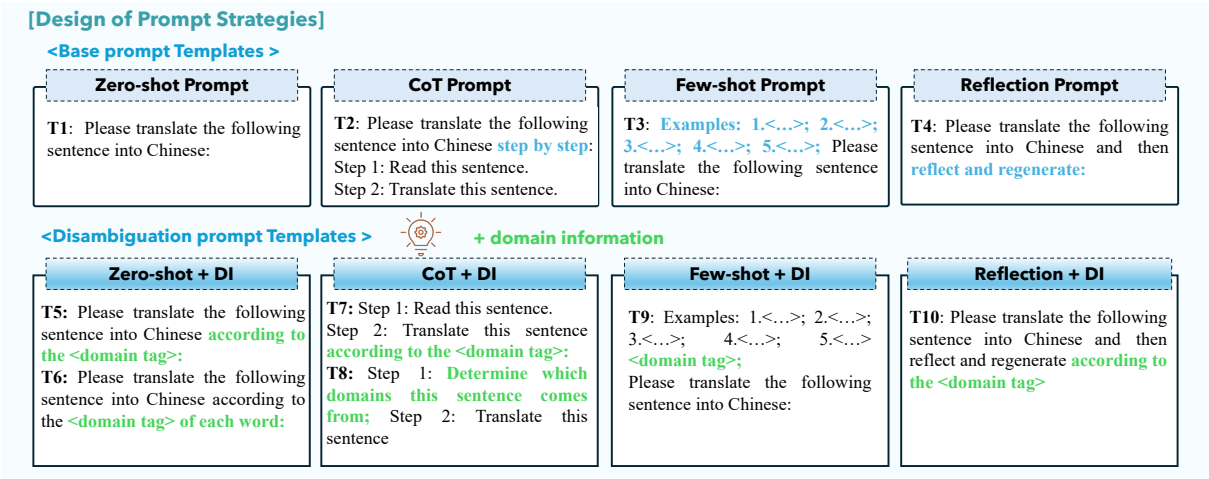

Figure 4: Design of Prompt Strategies. Light blue text represents for the specific information in each prompt strategy. Light green text represents for the specific information of disambiguation prompt strategies.

Chain-of-thought (CoT) prompts LLMs to reason about the input before generating an output (Wei et al., 2022). **(3) Few-shot**: this prompt supplies an LLM with task-specific examples before querying it (Brown, 2020). **(4) Reflection** (Shinn et al., 2024) further reflection on the generated translations yields new answers.

Disambiguation Prompt Strategies. In this work, our prompt includes 1) instructions to perform the task such as "*Please translate the following sentence into <target language>*" (*i.e.*,, T1), and 2) domain information such as domain tag. As shown in Figure 4, our disambiguation prompts strategies as following:

(1) Zero-shot + domain information: This strategies contain sentence-level and word-level: 1) Template 5 (T5): This template mainly utilize the domain information from the sentence-level domain tag base on the sentence-level MDT (Kobus et al., 2017). 1) Template 6 (T6): This template further utilize the domain information of each word base on the word-level MDT (Jiang et al., 2020). We aim to evaluate whether fine-grained domain information can disambiguate and improve the capability of LLMs' understanding.

(2) CoT + domain information: We also test whether CoT prompting could improve LLMs' performance by utilizing reasoning-based steps for quality evaluation, Template 4 "*Please translate the following sentence into <target language> step by step: Step 1: read this sentence. Step 2: translate this sentence.*" Moreover, we design two disambiguation prompting by devising Template 2: 1) Template 7: In this prompt, we give domain tag in step 2. This template further utilize domain information in reasoning ; 2) Template 8: In this prompt, we ask LLMs to automatically discriminate which domain the source sentence comes from in step 1.

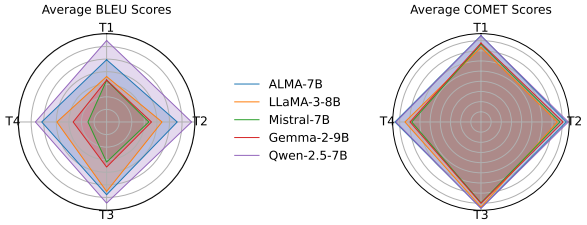

Figure 5: The comparison of different LLMs on the English-to-Chinese translation task (T1–T4) in terms of average BLEU and COMET scores.

(3) Few-shot + domain information: We randomly retrieves 5-shot examples from the training datastore and use these examples for translation, this prompt is Template 3. To further integrate domain information, we add domain tags to each example, enhancing LLM's ability to perceive domain as Template 9.

(4) Reflection + domain information: Reflection encourages LLMs to review and refine its responses for improved accuracy and coherence (Shinn et al., 2024). After reflecting on its initial output, the large model regenerates the translation as Template 4. We further enhance this process by incorporating domain information, encouraging the model to produce domain-specific translation results, as shown in Figure 4 Template 10.

2.3 Model Comparison and Selection

In order to achieve more accurate and cost-effective replication, we are using the popular open-source model available at present. Our model selection can be divided into the following three categories: **(1) Open-source**: we select LLama-3-8B (Grattafiori et al., 2024), Mistral-7B (Jiang et al., 2024), Gemma-2-9b (Team et al., 2024), and Qwen-2.5-7B which was specifically tested on a diverse set of 12 languages and showed impressive multilingual capabilities (Bai et al., 2023). **(2) LLM-based translation model**: ALMA-7B fine-tuned in Llama-3-7B with translation instructions (Xu et al., 2024). For all 5 selected models, we use the instruction-tuned version, i.e., the chat model, for zero-shot, CoT and few-shot inference. As shown in the Figure 5, Qwen-2.5-7B achieve the best performance on the English-Chinese MDT; therefore, we selected it as the base model for subsequent in-depth analysis in the section 3.1. **(3) NMT**: NLLB (Team et al., 2022) is a multilingual translation model developed by Meta AI, supporting 200 languages. In addition, to demonstrate the performance on a larger-scale model, we also compare with Qwen-2.5-14B. The specific results are shown in Appendix C Table 9 and 10.

2.4 Evaluation Metrics

Translation Quality. We adopt two widely-used metrics: SacreBLEU (Post, 2018), a n-gram matching-based metric, and the `wmt22-comet-da` model is used to generate the COMET[5] scores, the scope is 0-1, for convenience, we multiply the comet score by 100 in our experiments. In particular, we use the paired bootstrap resampling methods (Koehn, 2004) for the statistical significance test.

Disambiguation Accuracy. To evaluate the disambiguation ability of LLMs in MDT, we propose a metric based on the ambiguous vocabulary. Specifically, we identify all ambiguous source-language words in the test set and denote the total number of such instances as n. Among them, we count m instances where the words are correctly translated according to their domain-specific meanings. We define disambiguation accuracy as m/n, which reflects how effectively an LLM resolves lexical ambiguity across domains. For example, in the science domain, the word *"power"* should be translated as "能量" (*energy*) rather than "权力" (*authority*).

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

GPT-4o-mini Evaluator. Previous research (Qian et al., 2024) has shown that using GPT for translation quality evaluation is a feasible research approach. Therefore, we design a prompt to evaluate the disambiguation capability of LLMs using GPT-4o-mini[6].

3 Evaluation Experiments

In this section, we conduct an in-depth investigation of the three research questions (**RQs**) introduced in the Section 1 through experiments on English-Chinese translation.

Evaluation and Training. All our experiments were run using 1 × NVIDIA V100 32G, for different LLM variants. We use vLLM[7] (Kwon et al., 2023) to save inference time. We keep the parameters consistent with those used in previous work (Qian et al., 2024). For training, we use the Qwen-2.5-7B as base model for supervised fine-tuning base on the LLaMAFactory framework[8]. The Details of the training procedure parameters are provided in Appendix B.

3.1 Main Results

As shown in the Table 1, incorporating domain information proves effective for the vast majority of prompt strategies, indicating that the disambiguation prompt strategies play a positive role. However, we also observed several noteworthy phenomena, which we further analyze in detail below:

Finding 1: **Adding domain information does not improve performance across all base prompt strategies on average score.**

Analysis and Case Study for Finding 1. As shown in Table 2, for case 1, T6 adopts a word-based domain information translation strategy, focusing on lexical accuracy. As a result, the translations often exhibit clear word-to-word alignments, such as "a" → "一个" and "basin" → "盆子". T8 automatically determines the domain of the sentence and then translates accordingly. This strategy may lead to translation errors (*i.e.,* "contagion" → "传播") if the domain is Economic domain.

[6]The specific prompt template for GPT-4o-mini Evaluator in Appendix E

[7]https://github.com/vllm-project/vllm

[8]*https://github.com/hiyouga/LLaMA-Factory

Stragies		English-to-Chinese					AVG
		Education	Laws	News	Science	Spoken	
Zero-shot	T1	33.14 / 88.10	50.82 / 88.94	30.04 / 84.51	28.76 / 84.82	19.20 / 77.00	32.39 / 84.67
	T5	33.46 / 88.21	51.39 / 89.20	30.36 / 84.92	28.78 / 86.13	20.89 / 77.46	**32.98 / 85.18**
	T5-T1	+ 0.32 / + 0.11	+ 0.57 / + 0.26	+ 0.32 / + 0.41	+ 0.02 / + 1.31	+ 1.69 / + 0.46	+ 0.59 / + 0.51
	T6	32.64 / 87.84	50.10 / 88.29	30.10 / 84.25	27.99 / 85.50	18.40 / 75.06	31.85 / 84.19
	T6-T1	- 0.50 / - 0.26	- 0.72 / - 0.65	+ 0.06 / - 0.26	- 0.77 / + 0.68	- 0.80 / - 1.94	- 0.54 / - 0.48
CoT	T2	34.02 / 88.06	51.19 / 89.60	30.51 / 84.91	28.82 / 85.91	22.45 / 79.31	33.40 / 85.56
	T7	34.50 / 88.09	52.09 / 90.15	31.00 / 85.15	28.97 / 86.05	23.47 / 80.88	**33.99 / 86.06**
	T7-T2	+ 0.48 / + 0.03	+ 0.9 / + 0.55	+ 0.49 / + 0.24	+ 0.15 / + 0.14	+ 1.02 / + 1.57	+ 0.59 / + 0.50
	T8	33.56 / 88.22	50.39 / 88.79	30.15 / 84.88	28.95 / 86.05	22.02 / 79.01	32.61 / 85.79
	T8-T2	- 0.46 /+ 0.16	- 0.80 / -0.81	-0.36 / -0.03	+ 0.13 / + 0.14	-0.43 / - 0.30	- 0.79 / + 0.23
Few-shot	T3	34.17 / 88.17	50.48 / 89.22	29.91 / 84.66	28.33 / 85.64	18.44 / 77.82	**32.27 / 85.10**
	T9	33.63 / 88.03	50.46 / 89.49	29.76 / 84.68	27.94 / 85.89	18.05 / 77.32	31.97 / 85.08
	T9-T3	- 0.54 / - 0.14	- 0.02 /+ 0.27	- 0.15 / + 0.02	- 0.39 / + 0.25	- 0.39 / - 0.50	- 0.30 / - 0.02
Reflection	T4	26.75 / 86.06	47.77 / 87.76	26.16 / 82.71	25.90 / 84.03	17.01 / 76.05	28.72 / 83.32
	T10	32.80 / 87.83	50.61 / 89.16	30.24 / 84.57	28.60 / 85.68	22.20 / 79.40	**32.89 / 85.33**
	T10-T4	+ 6.05 / + 1.77	+ 2.84 / + 1.40	+ 4.08 / + 1.86	+ 2.70 / + 1.65	+ 5.19 / + 3.35	+ 4.17 / + 2.01
Stragies		**Chinese-to-English**					**AVG**
		Education	**Laws**	**News**	**Science**	**Spoken**	
Zero-shot	T1	22.19 / 83.05	36.03 / 83.48	17.63 / 80.31	16.52 / 81.36	10.39 / 64.43	20.55 / 78.53
	T5	26.61 / 84.02	34.00 / 83.37	18.35 / 80.94	17.68 / 81.86	11.07 / 64.97	**21.94 / 79.83**
	T5-T1	+ 4.42 / + 0.97	- 2.03 / - 0.11	+ 0.72 / + 0.63	+ 1.16 / + 0.50	+ 0.68 / + 0.54	+ 1.39 / + 1.30
	T6	25.05 / 83.44	33.42 / 82.67	16.20 / 78.38	16.88 / 80.39	10.12 / 64.30	20.33 / 77.84
	T6-T1	+ 2.86 / + 0.39	- 2.61 / - 0.81	- 1.43 / - 1.93	+ 0.36 / - 0.97	- 0.27 / - 0.13	- 0.22 / - 0.69
CoT	T2	26.65 / 84.17	33.49 / 83.44	17.82 / 80.60	18.12 / 81.82	13.87 / 70.05	21.99 / 80.02
	T7	26.63 / 84.91	34.46 / 83.72	18.08 / 80.29	18.82 / 81.86	14.59 / 71.97	**22.52 / 80.55**
	T7-T2	- 0.02 / + 0.74	+ 0.97 / + 0.28	+ 0.26 / - 0.31	+ 0.70 / + 0.04	+ 0.72 / + 1.92	+ 0.53 / + 0.53
	T8	26.38 / 84.23	33.88 / 83.54	17.89 / 80.72	18.18 / 81.90	13.25 / 69.68	21.92 / 79.99
	T8-T2	- 0.27 / + 0.06	+ 0.39 / + 0.10	+ 0.07 / + 0.12	+ 0.06 / + 0.08	- 0.62 / - 0.37	- 0.07 / - 0.03
Few-shot	T3	26.36 / 84.05	32.08 / 83.44	18.67 / 80.69	18.26 / 81.80	12.30 / 68.91	21.53 / 79.78
	T9	27.11 / 83.96	32.11 / 83.56	18.68 / 80.80	18.33 / 81.91	12.97 / 69.66	**21.84 / 79.98**
	T9-T3	+ 0.75 / - 0.09	+ 0.03 / + 0.12	+ 0.01 / + 0.11	+ 0.07 / + 0.11	+ 0.67 / + 0.75	+ 0.31 / + 0.20
Reflection	T4	15.55 / 78.96	28.33 / 80.06	16.07 / 79.35	15.15 / 79.86	10.11 / 65.50	17.04 / 76.75
	T10	24.33 / 83.69	34.25 / 83.77	17.72 / 80.68	16.84 / 81.87	13.52 / 66.30	**21.33 / 79.26**
	T10-T4	+ 8.78 / + 4.73	+ 5.92 / + 3.71	+ 1.65 / + 1.33	+ 1.69 / + 2.01	+ 3.41 / + 0.80	+ 4.29 / + 2.51

Table 1: BLEU and COMET scores on the English-Chinese translation task for T1-T10 with Qwen-2.5-7B. We **bold** the best performance results in each strategy (*hereinafter the same*). Orange text stands for the disambiguation prompting templates. " / " represents for the "BLEU / COMET". Blue text represents for the improvement and red text represents for the decrease (*hereinafter the same*).

Domain	Education
SRC	He washed his hands in a basin.
REF	他在盆里洗了手。
T1	他用盆洗了手。
T6	他用一个盆子洗了手。
Domain	**News**
SRC	Is there a suicide contagion on Wall Street?
REF	金融行业沦为自杀高发行业?
T2	华尔街存在自杀高发吗?
T8	华尔街是否存在自杀传播?

Table 2: Three cases illustrate the phenomenon of decreased average scores for T6 and T8.

Finding 2: **Apart from Reflection, adding domain information to other strategies yields inconsistent improvements across domains, even in the best-performing approach on average, CoT with domain information.**

Analysis for Finding 2. As shown in the Table 1, for both English-to-Chinese and Chinese-to-English translation direction, CoT combined with domain information (*i.e.,* T7) achieves the highest average BLEU and COMET scores, reaching "33.99 / 86.06" and "22.52 / 80.55", respectively. This indicates that the reasoning-based approach of LLMs can generate more accurate translations across multiple domains. Notably, the Reflection achieves consistent improvements across all domains when domain information is incorporated, suggesting that it effectively leverages domain knowledge during the reasoning process. In contrast, other strategies do not show consistent gains with domain information, which we hypothesize is due to their limited ability to enhance disambiguation performance.

Finding 3: **Different domains exhibit varying degrees of sensitivity to prompt templates.**

Domain	Laws
SRC	Chapter III Fundamental Rights and Duties of the Residents
REF	第三章 居民的基本权利和义务
SRC	Chapter IX Supplementary Provisions
REF	第九章附则

Table 3: Two cases illustrate the specialized domain terminology and distinct textual styles.

Analysis and Case Study for Finding 3. For Chinese-to-Chinese, in the zero-shot setting, the Spoken domain sees notable gains from T1 to T5, with BLEU increasing by 1.69 and COMET by 0.46, while the Science domain under the CoT strategy shows minimal change from T2 to T7, with BLEU increasing by only 0.15 and COMET by 0.14. In contrast, the Reflection strategy, comparing T4 and T10, achieves consistent and substantial improvements across all domains. For example, in the Education domain, BLEU increases by 6.05 and COMET by 1.77, this is due to the presence of more prominent domain features, such as specialized terminology and distinct textual styles, as shown in Table 3.

Finding 4: **In some domains, BLEU improves while COMET decreases, indicating that these metrics fail to adequately reflect the model's ability to handle ambiguity in MDT.**

Domain	News
SRC	It's clear he doesn't have any power.
REF	他显然没有任何权力。
T6	显然没有任何力量
T8	他显然没有任何权利。

Table 4: One case illustrate the phenomenon of BLEU and COMET scores are not inconsistent for T6 and T8.

Analysis and Case Study for Finding 4. For the English-to-Chinese translation direction, we found that BLEU and COMET scores exhibit divergent trends in the News domain. To illustrate this phenomenon, we present a case where the English word "*power*"—which can mean either "权势" (*authority*) or "力量" (*strength*)-demonstrates lexical ambiguity, as shown in Table 4. This ambiguity can lead to discrepancies in evaluation results when using BLEU and COMET, as each metric may favor different reference choices.

In summary, the aforementioned interesting findings further demonstrate the necessity of explicitly designing prompt templates to reveal and study the disambiguation capabilities of LLMs in MDT. In addition, we conducted experiments on multi-domain datasets for German–English, Japanese–English, and Korean–English translation directions[9]. The overall trends are consistent with those observed in the English–Chinese experiments, which supports the validity and rationality of our proposed research motivation. Detailed results are provided in Appendix D.

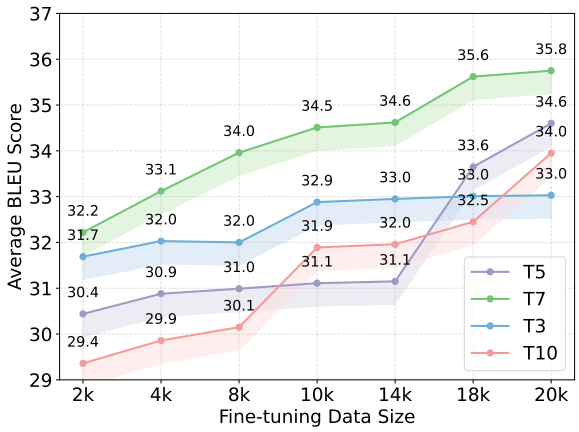

Figure 6: The average BLEU scores on the English-to-Chinese translation task with different fine-tuning data scales. The x-axis represents the amount of fine-tuning data selected from each domain.

3.2 Fine-tuning Resluts

Based on the experimental results in Table 1, we further fine-tune the prompt strategies that benefit from domain information on Qwen-2.5-7B. The specific fine-tuned results are shown in the Figure 6, with the increase in fine-tuning data, all prompt strategies exhibit improved average BLEU scores. Notably, T5 shows the greatest improvement, which further highlights the effectiveness of our proposed prompt design.

3.3 Disambiguation Performance

As shown in the Table 5, we further analyze the disambiguation performance of multiple disambiguation strategies, contain the following conclusion:

Conclusion 1 (⇒ *Finding 1*): The disambiguation performance of the templates also improves in consistency with adding domain information.

[9]Japanese-English and Korean-English from the Flores-101 data set (NLLB Team et al., 2024). We give a detailed data processing in the Appendix A

Strategies		English-to-Chinese					AVG
		Education	Laws	News	Science	Spoken	
Zero-shot	T1	39.68	40.85	46.89	36.98	42.88	41.46
	T5	42.56$_{+2.88}$	44.96$_{+4.11}$	47.69$_{+0.80}$	44.12$_{+7.14}$	43.65$_{+0.77}$	**44.60**$_{+3.14}$
	T6	36.36$_{-3.32}$	38.19$_{-2.66}$	45.11$_{-1.78}$	35.20$_{-1.78}$	40.56$_{-2.32}$	39.08$_{-2.38}$
CoT	T2	44.60	45.97	47.22	45.14	44.90	45.57
	T7	45.04$_{+0.44}$	46.98$_{+1.01}$	48.27$_{+1.05}$	46.05$_{+0.91}$	45.98$_{+1.08}$	**46.46**$_{+0.89}$
	T8	36.25$_{-8.35}$	38.69$_{-7.28}$	38.94$_{-8.28}$	30.10$_{-15.04}$	39.50$_{-5.40}$	36.70$_{-8.87}$
Few-shot	T3	40.55	41.60	48.11	37.25	43.00	42.10
	T9	34.58$_{-5.97}$	39.26$_{-2.34}$	42.30$_{-5.81}$	35.44$_{-1.81}$	40.87$_{-2.13}$	38.49$_{-3.61}$
Reflection	T4	43.56	42.07	47.85	39.60	44.09	43.43
	T10	45.02$_{+1.46}$	43.05$_{+0.98}$	47.28$_{-0.57}$	40.15$_{+0.55}$	45.60$_{+1.51}$	**44.22**$_{+0.79}$

Table 5: Disambiguation accuracy scores (%) on English-to-Chinese translation task for T1–T10 with Qwen-2.5-7B.

Conclusion 2 (⇒ *Finding 2*): The reason for the consistent improvement in translation performance under the reflection strategy lies in its ability to disambiguate the target translations.

Conclusion 3 (⇒ *Finding 3*): In domains with stronger features (*i.e.,* Law and Science domains), the improvement in disambiguation accuracy is also greater. This further demonstrates that using this metric in these domains can better assess translation performance.

Conclusion 4 (⇒ *Finding 4*): Disambiguation accuracy reflects improvements or declines in consistency, thereby avoiding inconsistencies in the increases or decreases of BLEU and COMET scores.

Overall, the proposed disambiguation accuracy further demonstrates the effectiveness of the prompt strategies and corroborates the findings discussed above. The detailed disambiguation accuracy results for German-English and GPT-4o-mini are provided in Appendix E, respectively.

4 Related work

Multi-domain Translation. MDT seeks to design a unified NMT model to translate texts across various domains, which can be divided into sentence-level (Kobus et al., 2017; Britz et al., 2017; Tars and Fishel, 2018; Aharoni and Goldberg, 2020) and word-level (Zeng et al., 2018; Su et al., 2021; Jiang et al., 2020; Lai et al., 2022; Zhang et al., 2021; Man et al., 2024b) domain representation learning. Recently, some researchers have explored the MDT on LLMs (Hu et al., 2024b; Zheng et al., 2024). These methods based on conventional encoder-decoder framework. However, we aim to explore the performance of disambiguate when utilizing the disambiguation prompt strategies in LLMs.

Disambiguation Evaluation for Translation. Ambiguity has long been a central challenge in machine translation, with numerous studies conducting evaluations in general domains (Campolungo et al., 2022; Maheshwari et al., 2024; Martelli et al., 2025; Hu et al., 2024a). In addition, some work has attempted to improve disambiguation by incorporating domain-specific dictionaries through constraint-based translation (Song et al., 2019; Chen et al., 2021; Zhang et al., 2023b; Baek et al., 2023). However, our approach fundamentally differs in that it does not rely on external constraint resources such as dictionaries. Instead, we focus on systematically evaluating and enhancing the disambiguation capabilities of LLMs.

LLMs for Translation. These work can be broadly divided into two main categories. The first category focuses on leveraging prompting techniques (Vilar et al., 2022; Jiao et al., 2023b; Zhang et al., 2023a; Moslem et al., 2023; He et al., 2024) to enhance and analyze the performance of machine translation using LLMs. The second category focuses on fine-tuning LLMs to improve their performance in downstream NLP tasks (Xu et al., 2023; Jiao et al., 2023a; Zeng et al., 2023). Our key contribution is identifying essential MDT disambiguation information for LLMs and designing prompt strategies.

5 Conclusion

In this work, we propose **DMDTEval**, a systematic benchmark for evaluating the disambiguation capabilities of LLMs in MDT. We construct a dedicated ambiguous word dataset, explore diverse prompting strategies, and evaluate five leading open-source LLMs across four language pairs and thirteen domains. Our analysis reveals key challenges in MDT disambiguation and provides actionable insights for improving domain-aware translation. In future work, we plan to develop improved methods and mechanism for disambiguation building upon the proposed ambiguous word dataset.

Limitations

Multi-domain bilingual parallel corpora are scarce and difficult to obtain at scale, which poses a fundamental challenge for research on MDT. Although the datasets used in our study are widely adopted in previous work and cover a broad range of domains, we acknowledge that they may contain noise, inconsistencies, or domain overlaps that could affect evaluation outcomes. Furthermore, the provenance and annotation quality of some datasets are not always transparent or verifiable, which may introduce bias into the model's disambiguation assessment. Another limitation lies in the uneven distribution of domain data across different language pairs. While we focus on English–Chinese and German–English due to our linguistic expertise, other language pairs (*e.g.,* Japanese–English and Korean–English) are only used for evaluation purposes and lack the same level of manual verification and ambiguous word coverage. In future work, we plan to address these limitations by focusing on high-quality data collection and annotation.

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Yujin Baek, Koanho Lee, Dayeon Ki, Hyoung-Gyu Lee, Cheonbok Park, and Jaegul Choo. 2023. Towards accurate translation via semantically appropriate application of lexical constraints. *arXiv preprint arXiv:2306.12089*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Denny Britz, Quoc Le, and Reid Pryzant. 2017. Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126, Copenhagen, Denmark. Association for Computational Linguistics.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Niccolò Campolungo, Federico Martelli, Francesco Saina, Roberto Navigli, et al. 2022. Dibimt: A novel benchmark for measuring word sense disambiguation biases in machine translation. In *PROCEEDINGS OF THE CONFERENCE-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. MEETING*, pages 4331–4352.

Guanhua Chen, Yun Chen, Yong Wang, and Victor OK Li. 2021. Lexical-constraint-aware neural machine translation via data augmentation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3587–3593.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128.

Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Zhe Xu, Yao Hu, Jian Wu, and Zuozhu Liu. 2025. Mt-r1-zero: Advancing llm-based machine translation via r1-zero-like reinforcement learning. *arXiv preprint arXiv:2504.10160*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer,

Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai

Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Tianxiang Hu, Pei Zhang, Baosong Yang, Jun Xie, Derek Wong, and Rui Wang. 2024a. Large language model for multi-domain translation: Benchmarking and domain cot fine-tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5726–5746.

Tianxiang Hu, Pei Zhang, Baosong Yang, Jun Xie, Derek F. Wong, and Rui Wang. 2024b. Large language model for multi-domain translation: Benchmarking and domain CoT fine-tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5726–5746, Miami, Florida, USA. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Haoming Jiang, Chen Liang, Chong Wang, and Tuo Zhao. 2020. Multi-domain neural machine translation with word-level adaptive layer-wise domain mixing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1823–1834, Online. Association for Computational Linguistics.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023a. Parrot: Translating during chat using large language models tuned with human translation and feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15009–15020.

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

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria. INCOMA Ltd.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Wen Lai, Jindřich Libovický, and Alexander Fraser. 2022. Improving both domain robustness and domain adaptability in machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5191–5204, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Frederick Liu, Han Lu, and Graham Neubig. 2018. Handling homographs 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 1336–1345, New Orleans, Louisiana. Association for Computational Linguistics.

Ayush Maheshwari, Preethi Jyothi, and Ganesh Ramakrishnan. 2024. Dictdis: Dictionary constrained disambiguation for improved nmt. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10991–11004.

Zhibo Man, Kaiyu Huang, Yujie Zhang, Yuanmeng Chen, Yufeng Chen, and Jinan Xu. 2024a. ICL: Iterative continual learning for multi-domain neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7732–7743, Miami, Florida, USA. Association for Computational Linguistics.

Zhibo Man, Zengcheng Huang, Yujie Zhang, Yu Li, Yuanmeng Chen, Yufeng Chen, and Jinan Xu. 2023. Wdsrl: Multi-domain neural machine translation with word-level domain-sensitive representation learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:577–590.

Zhibo Man, Zengcheng Huang, Yujie Zhang, Yu Li, Yuanmeng Chen, Yufeng Chen, and Jinan Xu. 2024b. Wdsrl: Multi-domain neural machine translation with word-level domain-sensitive representation learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:577–590.

Zhibo Man, Yujie Zhang, Yu Li, Yuanmeng Chen, Yufeng Chen, and Jinan Xu. 2024c. An ensemble strategy with gradient conflict for multi-domain neural machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(2):1–22.

Federico Martelli, Stefano Perrella, Niccolò Campolungo, Tina Munda, Svetla Koeva, Carole Tiberius, and Roberto Navigli. 2025. Dibimt: A gold evaluation benchmark for studying lexical ambiguity in machine translation. *Computational Linguistics*, pages 1–71.

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

NLLB Team, 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, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

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

Shenbin Qian, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orasan, Tharindu Ranasinghe, and Fred Blain. 2024. What do large language models need for machine translation evaluation? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3660–3674, Miami, Florida, USA. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing nmt with pre-specified translation. *arXiv preprint arXiv:1904.09107*.

Jinsong Su, Jiali Zeng, Jun Xie, Huating Wen, Yongjing Yin, and Yang Liu. 2021. Exploring discriminative word-level domain contexts for multi-domain neural machine translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1530–1545.

Sander Tars and Mark Fishel. 2018. Multi-domain neural machine translation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 279–288, Alicante, Spain.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Kowalevski, Paul Koura, Vitaliy Liptchinsky, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Daniel Youngblood, Omar Zaidan, et al. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, Yi Lu, Shuo Li, Yiming Wang, and Longyue Wang. 2014. UM-corpus: A large English-Chinese parallel corpus for statistical machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1837–1842, Reykjavik, Iceland. European Language Resources Association (ELRA).

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance. *arXiv preprint arXiv:2211.09102*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. *Preprint*, arXiv:2309.11674.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. Tim: Teaching large language models to translate with comparison. *arXiv preprint arXiv:2307.04408*.

Jiali Zeng, Jinsong Su, Huating Wen, Yang Liu, Jun Xie, Yongjing Yin, and Jianqiang Zhao. 2018. Multi-domain neural machine translation with word-level domain context discrimination. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 447–457, Brussels, Belgium. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Jinpeng Zhang, Nini Xiao, Ke Wang, Chuanqi Dong, Xiangyu Duan, Yuqi Zhang, and Min Zhang. 2023b. Disambiguated lexically constrained neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10583–10596.

Shiqi Zhang, Yan Liu, Deyi Xiong, Pei Zhang, and Boxing Chen. 2021. Domain-aware self-attention for multi-domain neural machine translation. *Proc. Interspeech 2021*, pages 2047–2051.

Jiawei Zheng, Hanghai Hong, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning large language models for domain-specific machine translation. *arXiv preprint arXiv:2402.15061*.

A Dataset

As shown in Table 6, the training and testing data sizes for the English-Chinese and German-English datasets are presented. Besides these two language pairs, we use data sets from FLORES[10] (Team et al., 2022), selecting Japanese-English and Korean-English. The test set consists of 1,012 sentences covering three domains: Wikinews, Wikibooks, and Wikiyago, referred to in paper as the news, book, and travel domains. After domain-wise splitting, the data for these three domains consist of 341, 351, and 321 sentences, respectively. Additionally, regarding the scoring of alignment quality, we observe that our human annotated results achieve higher accuracy, as shown in Table 7.

B Evaluation and Training Details

Specifically, we chose the default hyperparameter settings in vLLM for all our experiments, *i.e.,* 0.8 as temperature 4, 0.95 for top_p. The input sequence length was chosen as 1024 for zero-shot and CoT inference and 3000 for few-shot inference. For the training procedure, we utilize the LoRA (Hu et al., 2021) to fine-tune the Qwen-2.5-7B. The hyper-parameters for supervised fine-tuning are listed in Table 8.

[10]https://huggingface.co/datasets/facebook/flores

English-Chinese					
Train set	Edu	Laws	News	Sci	Spo
	444K	207K	443K	263K	210K
Test set	Edu	Laws	News	Sci	Spo
	790	456	1500	503	455
German-English					
Train set	IT	Kor	Laws	Med	Sub
	211K	16K	434K	233K	470K
Test set	IT	Kor	Laws	Med	Sub
	2000	2000	2000	2000	2000

Table 6: The statistics of multi-domain translation data sets. Edu represents for the Education domain, Sci represents for the Science domain, Spo represents for the spoken domain, and Sub represents for the subtitles domain.

Domains	**C**	**Pc**	**I**
Education	89%	9%	2%
Laws	95%	4%	1%
News	84%	14%	2%
Science	87%	10%	3%
Spoken	82%	16%	2%

Table 7: C represents for the Correct label, Pc represents for the Partially correct label, and I represents for the Incorrect, respectively.

C Detailed Results on LLMs

As shown in Table 9 and Table 10, the two tables present the detailed results of our main experiments, including those from the neural machine translation model NLLB, the larger-scale model Qwen-2.5-14B, as well as the results obtained using GPT-4o-mini.

D Results on Other Language Pairs

As shown in Tables 11, 12, and 13, we provide detailed experimental results for German-English, Japanese-English, and Korean-English. The overall trends are consistent with those observed in English-Chinese, further demonstrating the effectiveness of the disambiguation prompt strategies. For German-to-English, Table 11 shows that disambiguation prompting (T5–T10) consistently improves translation quality over the baseline prompts

Hyper-Parameter	Value
lora rank	8
learning rate	1e-5
train epoch	2
per_device_batchsize	1
warm up ratio	0.1
learning rate scheduler	cosine

Table 8: Hyper-parameters for supervised fine-tuning

(T1–T4) across all domains. Notably, gains are most prominent in technical domains like Medical and Laws. This highlights the effectiveness of our disambiguation strategy in enhancing both BLEU and COMET scores for German-to-English translation.

E Disambiguation Performance and GPT-4o-mini Evaluator

Disambiguation Accuracy. As shown in Table 14, we present the disambiguation accuracy for German-to-English, further demonstrating the robustness of our evaluation framework across multiple language pairs. This indicates that our proposed method is not limited by specific languages and can be effectively generalized to other language directions.

GPT-4o-mini Evaluator. We design a prompt to evaluate the disambiguation capability of LLMs using GPT-4o-mini. The specific prompt is: "*source sentence: < >, target sentence: < >, generate sentence: < >. Please find the ambiguous word pairs in the source language sentence and the target language sentence, and count the number of ambiguous word pairs. Refer to the above word pairs to further count the accuracy of disambiguation in the generated sentences.* ". We calculate the average accuracy across different templates with GPT-4o-mini in Figure 7 and 8. The consistency with Figure 7 and 8 further prove the effectiveness of disambiguation prompting.

		Education	Laws	News	Science	Spoken	AVG
Neural Machine Translation-based Methods							
NLLB	/	25.03 / 79.65	38.61 / 84.73	16.85 / 74.87	17.37 / 76.40	11.43 / 72.86	21.86 / 77.70
Translation-based LLMs							
ALMA-7B	T1	27.86 / 86.87	23.35 / 88.95	28.57 / 84.02	25.39 / 84.39	17.59 / 76.32	24.75 / 84.91
	T2	30.45 / 87.14	41.43 / 89.24	28.13 / 83.99	25.79 / 84.36	18.65 / 76.36	28.09 / 84.02
	T3	29.64 / 86.86	43.41 / 89.54	27.22 / 83.65	26.00 / 84.51	18.45 / 77.02	28.94 / 84.72
	T4	27.86 / 86.88	24.91 / 89.07	28.26 / 83.93	25.82 / 84.50	18.21 / 77.52	25.81 / 84.58
Open-source LLMs							
LLaMA-3-8B	T1	22.97 / 77.40	22.88 / 71.30	16.03 / 72.31	15.81 / 74.04	15.62 / 72.46	18.06 / 73.30
	T2	22.70 / 79.50	31.21 / 73.31	21.32 / 74.89	19.87 / 76.00	17.10 / 72.05	22.04 / 75.15
	T3	28.20 / 86.67	43.27 / 87.67	23.92 / 82.95	22.59 / 84.01	18.20 / 73.45	27.64 / 83.75
	T4	20.37 / 78.76	26.53 / 73.23	17.77 / 76.09	17.67 / 76.69	15.51 / 70.32	19.77 / 74.82
Mistral-7B	T1	14.86 / 77.96	26.01 / 79.96	16.22 / 77.40	15.68 / 78.56	10.21 / 68.69	16.60 / 76.51
	T2	19.04 / 81.53	24.10 / 79.76	15.71 / 77.90	15.09 / 80.34	10.08 / 68.88	16.80 / 77.68
	T3	18.22 / 82.54	26.12 / 82.88	17.01 / 79.63	16.21 / 80.97	11.23 / 69.04	15.96 / 79.61
	T4	10.99 / 74.26	7.38 / 66.07	6.27 / 67.15	7.42 / 70.10	8.03 / 65.11	7.42 / 68.14
Gemma-2-9B	T1	15.62 / 77.05	20.03 / 81.87	15.96 / 78.28	17.66 / 78.54	12.10 / 72.33	16.67 / 77.61
	T2	16.32 / 79.09	20.36 / 83.23	16.56 / 79.51	18.16 / 80.83	12.06 / 71.03	17.85 / 80.67
	T3	18.12 / 81.08	20.66 / 83.35	16.78 / 79.80	18.99 / 82.79	13.11 / 72.86	17.93 / 79.98
	T4	14.69 / 71.78	13.16 / 69.12	12.33 / 70.57	15.25 / 71.26	11.10 / 66.42	13.31 / 69.83
Qwen-2.5-7B	T1	33.14 / 88.10	50.82 / 88.94	30.04 / 84.51	28.76 / 84.82	19.20 / 77.00	32.39 / 84.67
	T2	34.02 / 88.06	51.19 / 89.60	30.51 / 84.91	28.82 / 85.91	22.45 / 79.31	33.40 / 85.56
	T3	34.17 / 88.17	50.48 / 89.22	29.91 / 84.66	28.33 / 85.64	18.44 / 77.12	32.27 / 84.96
	T4	26.75 / 86.06	47.77 / 87.76	26.16 / 82.71	25.90 / 84.03	17.01 / 76.05	28.72 / 83.32
Qwen-2.5-14B	T1	36.14 / 89.45	53.69 / 89.36	34.75 / 87.20	30.55 / 88.22	23.65 / 80.02	35.76 / 86.85
	T2	**37.90 / 89.65**	**53.87 / 89.98**	**35.14 / 87.72**	**31.04 / 88.80**	**23.51** / 80.00	**36.29 / 87.23**
	T3	35.77 / 88.12	52.94 / 89.12	34.58 / 86.87	30.23 / 88.01	23.55 / 79.49	35.41 / 86.32
	T4	37.19 / 89.82	53.16 / 89.25	34.64 / 87.58	30.82 / 88.03	23.42 / **80.05**	35.85 / 86.95

Table 9: BLEU and COMET scores on the English-to-Chinese translation task (T1-T4) with different open-source LLMs and NMT models. The best results are highlighted in **bold**.

English-to-Chinese						
	Education	**Laws**	**News**	**Science**	**Spoken**	**AVG**
T1	36.77 / 88.74	50.32 / 90.31	33.00 / 85.51	30.91 / 86.61	20.25 / 79.68	34.25/ 86.17
T5	36.90 / 88.79	50.52 / 90.39	33.19 / 85.60	30.93 / 86.77	21.55 / 80.01	**34.62 / 86.31**
T6	36.58 / 88.60	50.47 / 80.42	32.55 / 85.47	30.25 / 86.39	19.50 / 79.22	33.87 / 84.02
T2	37.22 / 88.15	49.81 / 90.19	32.68 / 85.45	31.08 / 86.24	22.56 / 80.97	34.67 / 86.20
T7	38.96 / 88.61	49.86 / 90.21	32.74 / 85.54	32.88 / 86.42	23.45 / 81.05	**35.58 / 86.37**
T8	34.22 / 80.23	45.36 / 85.63	30.26 / 83.00	31.11 / 85.01	22.09 / 79.65	32.61 / 82.70
T3	35.15 / 88.40	52.63 / 90.54	33.12 / 85.48	30.74 / 86.59	20.13 / 79.60	34.35 / 86.12
T9	36.99 / 89.32	52.62 / 90.69	33.25 / 86.55	31.45 / 87.62	22.11 / 80.87	**35.28 / 87.01**
T4	35.97 / 88.50	50.60 / 90.14	32.74 / 85.43	31.24 / 86.76	21.69 / 80.88	34.45 / 86.34
T10	36.71 / 89.12	50.86 / 90.33	33.16 / 86.19	31.42 / 86.01	22.17 / 80.89	**34.86 / 86.51**

Table 10: BLEU and COMET scores on the English-to-Chinese translation task for T1-T10 with **GPT-4o-mini**. We **bold** the best performance results in each strategy. Orange text stands for the disambiguation prompt strategies.

German-to-English						
	IT	**Koran**	**Laws**	**Medical**	**Subtitles**	**AVG**
T1	29.67 / 60.08	12.85 / 69.12	25.20 / 75.69	22.06 / 72.67	20.65 / 74.20	22.09 / 70.35
T5	30.37 / 72.03	15.10 / 70.26	31.05 / 80.33	32.26 / 76.98	25.42 / 74.86	**26.84 / 74.89**
T6	23.54 / 67.82	14.64 / 69.79	28.42 / 78.66	32.63 / 77.63	18.74 / 70.77	23.59 / 72.93
T2	33.56 / 80.28	15.50 / 71.38	32.98 / 82.87	36.41 / 81.64	26.40 / 76.95	**28.97 / 78.62**
T7	33.01 / 77.45	14.34 / 70.50	29.97 / 81.38	33.47 / 79.96	25.54 / 76.62	27.27 / 77.18
T8	30.91 / 72.82	14.60 / 70.32	29.59 / 79.93	31.86 / 77.69	25.05 / 76.13	26.40 / 75.38
T3	33.67 / 77.75	15.61 / 71.24	33.18 / 82.61	34.58 / 79.59	25.72 / 76.52	28.55 / 77.54
T9	33.34 / 77.50	15.78 / 71.46	33.90 / 82.71	34.30 / 79.61	25.82 / 76.51	**28.63 / 77.56**
T4	28.90 / 76.01	13.52 / 69.48	29.03 / 80.50	33.01 / 79.29	21.16 / 73.77	25.13 / 75.81
T10	32.14 / 77.74	15.41 / 71.15	30.97 / 82.03	33.34 / 79.45	25.48 / 75.99	**27.47 / 77.27**

Table 11: BLEU and COMET scores on the German-to-English translation task for T1-T10 with Qwen-2.5-7B. We **bold** the best performance results in each strategy. Orange text stands for the disambiguation prompting

Japanese-to-English				
	Book	**Travel**	**News**	**AVG**
T1	23.00 / 85.38	17.47 / 84.65	21.86 / 85.97	20.78 / 85.33
T5	23.66 / 85.96	17.87 / 85.02	22.34 / 86.29	**21.29 / 85.76**
T6	21.07 / 83.45	16.75 / 84.01	21.43 / 85.66	19.75 / 84.37
T2	24.36 / 86.36	18.50 / 85.20	23.90 / 87.49	22.25 / 86.35
T7	24.71 / 86.48	18.69 / 85.48	24.23 / 87.60	**22.54 / 86.52**
T8	22.14 / 83.75	16.85 / 84.10	20.10 / 85.01	19.70 / 84.29
T3	22.96 / 84.03	16.89 / 83.99	20.65 / 83.20	20.17 / 83.74
T9	23.20 / 84.76	17.56 / 84.79	21.94 / 86.01	**20.90 / 85.19**
T4	24.10 / 85.96	17.97 / 85.00	21.98 / 86.01	**21.35 / 85.66**
T10	22.80 / 85.10	17.12 / 84.33	21.01 / 85.46	20.31 / 84.96

Table 12: BLEU and COMET scores on the Japanese-to-English translation task for T1-T10 with Qwen-2.5-7B. We **bold** the best performance results in each strategy. Orange text stands for the disambiguation prompt strategies.

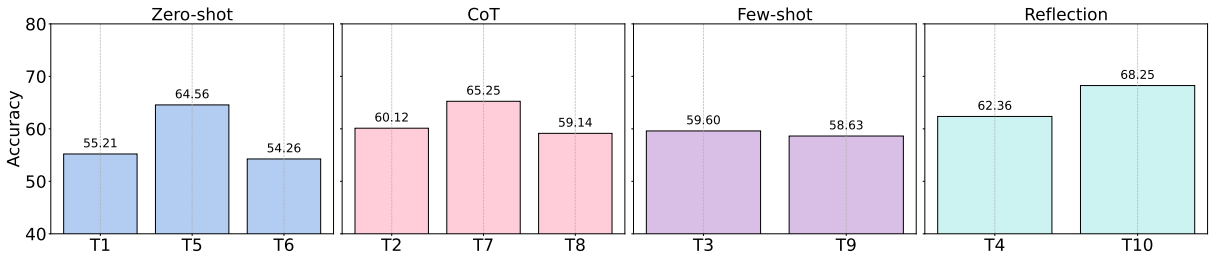

Figure 7: The accuracy of GPT-4o-mini Evaluator on the English-to-Chinese translation task.

	Korean-to-English			
	Book	**Travel**	**News**	**AVG**
T1	23.34 / 85.55	21.22 / 85.69	21.81 / 86.25	22.12 / 85.83
T5	24.05 / 86.15	22.64 / 85.78	22.77 / 86.10	**23.15 / 86.01**
T6	22.98 / 84.61	20.48 / 85.14	20.45 / 86.01	21.30 / 85.25
T2	23.98 / 85.70	21.58 / 85.74	22.60 / 86.43	22.72 / 85.96
T7	24.13 / 86.57	22.80 / 86.05	22.86 / 86.51	**23.26 / 86.38**
T8	23.01 / 85.09	21.17 / 85.44	21.65 / 86.08	21.94 / 85.54
T3	23.01 / 85.19	21.08 / 85.46	22.64 / 86.41	22.24 / 85.69
T9	23.58 / 85.67	21.85 / 85.70	22.80 / 86.57	**22.74 / 85.98**
T4	23.88 / 85.60	22.14 / 86.10	21.85 / 86.31	22.62 / 86.00
T10	23.98/ 85.62	22.45 / 86.13	22.07/ 86.50	**22.83 / 86.08**

Table 13: BLEU and COMET scores on the Korean-to-English translation task for T1-T10 with Qwen-2.5-7B. We **bold** the best performance results in each strategy. Orange text stands for the disambiguation prompt strategies

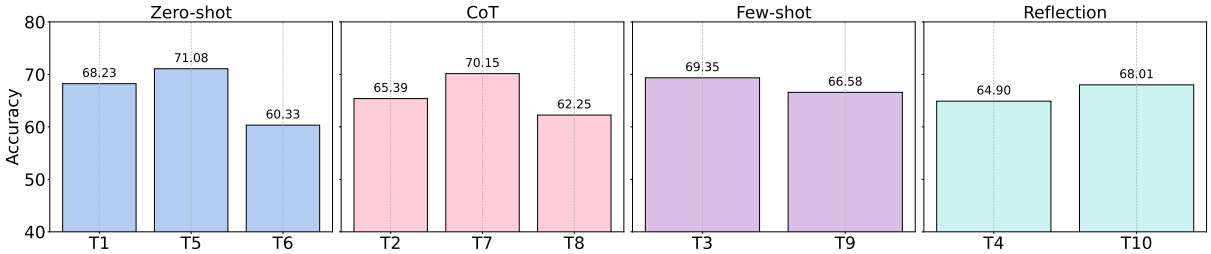

Figure 8: The accuracy of GPT-4o-mini Evaluator on the German-to-English translation task.

Strategies		German-to-English					AVG
		IT	**Koran**	**Laws**	**Medical**	**Subtitles**	
Zero-shot	T1	51.87	53.42	59.11	50.93	55.07	54.08
	T5	58.26$_{+6.39}$	57.89$_{+4.47}$	60.12$_{+1.01}$	57.34$_{+6.41}$	58.90$_{+3.83}$	**58.50**$_{+4.42}$
	T6	48.62$_{-3.25}$	52.15$_{-5.27}$	56.49$_{-3.63}$	48.80$_{-8.54}$	51.23$_{-7.67}$	51.46$_{-2.62}$
CoT	T2	55.89	58.07	60.35	56.88	57.31	57.70
	T7	58.02$_{+2.13}$	59.86$_{+1.79}$	61.17$_{+0.82}$	58.02$_{+1.14}$	58.67$_{+1.36}$	**59.15**$_{+1.45}$
	T8	44.99$_{-13.03}$	48.25$_{-10.62}$	47.15$_{-14.02}$	43.08$_{-14.94}$	49.19$_{-9.48}$	46.53$_{-11.17}$
Few-shot	T3	53.10	54.97	60.45	51.65	55.81	**55.20**
	T9	46.23$_{-6.87}$	50.78$_{-4.19}$	53.62$_{-6.83}$	48.84$_{-2.81}$	51.87$_{-3.94}$	50.27$_{-4.93}$
Reflection	T4	56.92	53.75	58.90	52.41	56.30	55.46
	T10	60.02$_{+3.10}$	57.81$_{+4.06}$	59.51$_{+0.61}$	55.89$_{+3.48}$	59.48$_{+3.18}$	**58.54**$_{+3.08}$

Table 14: Disambiguation accuracy scores (%) on the German-to-English translation task for T1-T10 with Qwen-2.5-7B. We **bold** the best performance results in each strategy.