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 prompting templates, and **(3)** we design precise disambiguation metrics, and study the efficacy of various prompting strategies on multiple state-of-the-art LLMs. 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.

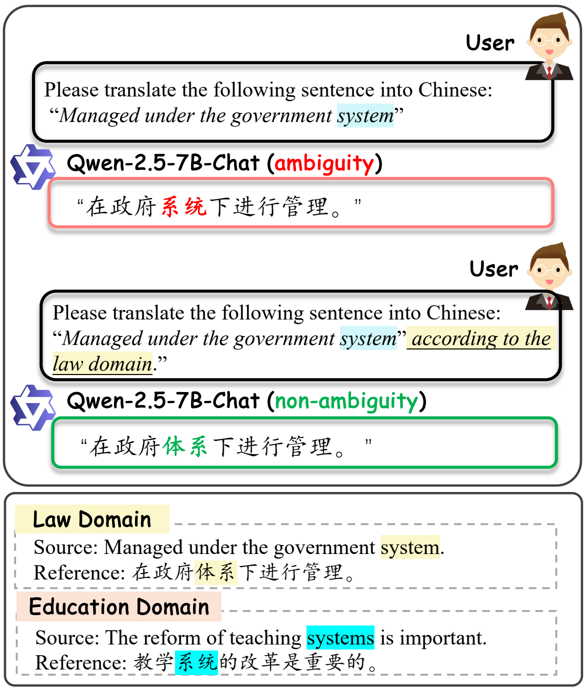

Figure 1: Our motivation. We prompt LLMs with domain label to disambiguate in Qwen-2.5-7B-Instruct.

1 Introduction

In recent years, Large Language Models (LLMs) have demonstrated increasing capabilities and applications across various Natural Language Processing (NLP) tasks (Zhao et al., 2023; Pan et al., 2023; Qin et al., 2024), benefiting from the ultra-large-scale training data and computation resources. Notably, LLMs achieving promising results in machine translation (MT) that demonstrate their potential in practical applications (Jiao et al., 2023b; Hendy et al., 2023; Wang et al., 2024; Feng et al., 2024; Qian et al., 2024).

However, the performance of LLMs in multi-domain translation (MDT) is less than satisfactory (Jiao et al., 2023b; Zheng et al., 2024; Hu et al., 2024b). This is because LLMs rely on extensive pre-training data, but multi-domain parallel training corpora are exceedingly scarce. This scarcity restricts the translation capabilities of LLMs, hindering their ability to effectively acquire cross-domain knowledge and leading to translation ambiguities. As shown in Figure 1, directly using LLMs for translation leading to word ambiguities. For instance, the term "system" which should be translated as "体系" (*framework*), may be mistranslated as "系统" (*the literal translation of system*), this presents a significant challenge for LLMs in multi-domain translation. However, when domain-specific information is provided, the translations become more accurate. The critical issue, therefore, is how to effectively utilize prompting to enhance the representation of domain-specific information. Recent studies have explored the performance of LLMs in multi-domain translation (Hu et al., 2024b) and investigated fine-tuning LLMs using domain-specific parallel corpora (Hu et al., 2024b; Zheng et al., 2024). However, three key research questions (RQ) remain unresolved in MDT:

- ♣ **RQ1**: *How to quantify the disambiguation ability of LLMs in multi-domain translation?*
- ♣ **RQ2**: *Can various prompting techniques help LLMs disambiguate in multi-domain translation?*
- ♣ **RQ3**: *What domain knowledge is essential for LLMs to achieve effective multi-domain translation?*

To answer and explore the aforementioned questions, we introduce a novel 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 prompting templates to evaluate the translation performance of prominent LLMs across multiple domains. **For RQ3**: We conduct extensive experiments and provide a detailed analysis and findings based on these experimental results.

In summary, the main contributions of our work can be summarized as follows:

- To the best of our knowledge, we are the first to construct a MDT test set with ambiguous word annotations, containing 10 domains across 2 language pairs: English-Chinese and German-English. We will open-source this dataset to facilitate research on the disambiguation of LLMs in MDT.
- We systematically explore various disambiguation prompting strategies, including zero-shot, chain-of-thought (CoT), few-shot, and reflection prompting, to evaluate multi-domain translation quality using 5 popular open-source LLMs.
- We investigate the types of domain knowledge required by LLMs to evaluate translations across 2 language pairs and 10 domains, focusing on sentence-level and word-level domain knowledge, domain-specific examples, and domain discrimination capabilities.

The rest of the paper is structured as follows: Section 2 discusses relevant work in MDT, while Section 3 introduces the overall evaluation framework. Section 4 describes the processing of test set construction. Section 5 describes the disambiguation prompting templates. Evaluation metrics are presented in Section 6. Section 7 concludes our study and outlines future directions.

2 Related work

Multi-domain Translation. Multi-domain translation seeks to design a unified NMT model to translate texts across various domains, which can be divided into sentence-level and word-level domain representation learning methods. Sentence-level strategy consists mainly of: Domain Tag (Kobus et al., 2017), Domain Discriminator (Britz et al., 2017), and Auto Clustering (Tars and Fishel, 2018; Aharoni and Goldberg, 2020). Word-level strategy: Zeng et al. (2018) and Su et al. (2021) design a word-level discriminator to learn the domain-specific representation of words. Previous work proposes domain proportion mechanism to improve the adaptability of each individual word domain (Jiang et al., 2020; Lai et al., 2022; Zhang et al., 2021; Man et al., 2023, 2024b,c). In summary, the above-mentioned traditional methods based on conventional encoder-decoder framework. However, we aim to explore the performance of disambiguate when utilizing the disambiguation prompting templates in LLMs.

LLMs for Machine Translation. LLMs for machine translation can be broadly divided into three main categories. The first category focuses on leveraging prompting techniques (Vilar et al., 2022; Jiao et al., 2023b; Zhang et al., 2023; Moslem et al., 2023; He et al., 2024) to enhance and analyze the performance of machine translation using LLMs, such as multilingual translation (Zhu et al., 2024), translation evaluation (Qian et al., 2024), low-resource translation (Moslem et al., 2023), document-level translation, multi-domain translation (Hu et al., 2024a), and etc. The second category focuses on fine-tuning LLMs to improve their performance in downstream NLP tasks. For example, ALMA (Xu et al., 2023) leverages monolingual data for low-resource languages, and ParroT (Jiao et al., 2023a), which uses human-written

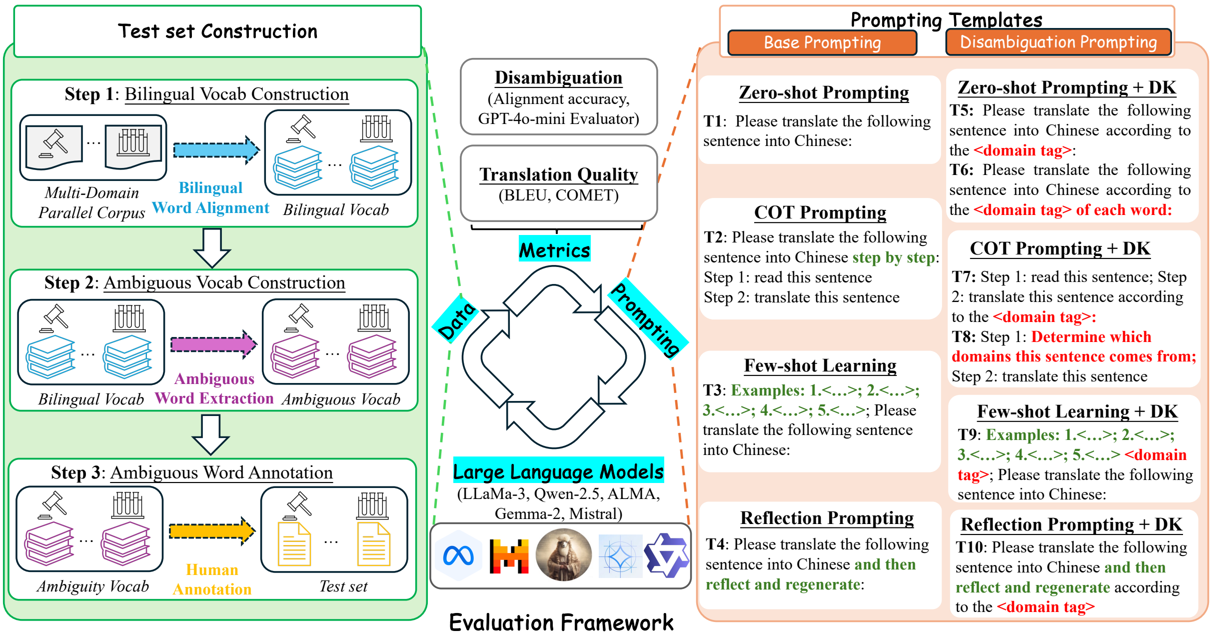

Figure 2: Our *DMDTEval* evaluation framework consists of Data, Prompting, LLMs, and Metrics. DK represents for the domain knowledge.

translations and feedback for high-quality results, have been developed. Additionally, TIM (Zeng et al., 2023) utilizes comparison data to guide models toward better translations. Our approach falls within the first category, building upon previous research (Zhu et al., 2024; Qian et al., 2024; Hu et al., 2024a). However, the key distinction lies in our focus on identifying the necessary information for MDT with LLMs for disambiguate, and further designing various disambiguation prompting mechanisms to enhance translation performance.

3 DMDTEval: Evaluation Framework

In our work, our goal includes (1) evaluating the influence of domain information in LLMs' translation, (2) and designing the metrics of word ambiguity. Therefore, we design a complete evaluation framework, as shown in Figure 2.

Data. 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 multi-domain machine translation test sets: UM-Corpus[1] (English-Chinese), including five domains: *Education, Law, News Science, and Subtitles* (Tian et al., 2014), and OPUS[2] (German-English), including five domains: *IT, Koran, Laws, Medical, and Subtitles* (Aharoni and Goldberg, 2020). We utilize the train set from these domains to obtain an ambiguity vocabulary, which is then used to annotate the test set for these domains. The overall statistics are listed in Table 1. The detailed construction process of the test set in Section 4.

[1]http://nlp2ct.cis.umac.mo/um-corpus/
[2]http://opus.nlpl.eu/

English-Chinese					
Train set	Edu	Laws	News	Sci	Sub
	444K	207K	443K	263K	220K
Test set	Edu	Laws	News	Sci	Sub
	790	456	1500	503	597
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 1: The statistics of multi-domain translation data sets. Edu represents for the Education domain, Sci represents for the Science domain, and Sub represents for the subtitles domain.

Prompting Templates. LLMs translate a source language into a target language with prompts. Therefore, designing an effective prompt is the key to unlocking the translation capabilities of LLMs. In our work, we evaluate impact of different base prompting strategies, including: Zero-shot learning, Chain-of-Thought (COT), Few-shot learning, and Reflection, and we further focus on design disambiguation prompting to improve translation performance of LLMs in Section 5: (1) **Zero-shot prompting** directly asks LLM to translate a source input into the target language (Liu et al., 2018). (2) **Chain-of-thought (CoT)** prompts LLMs to reason about the input before generating an output (Wei et al., 2022). (3) **Few-shot prompting** (i.e., *In-context learning*): this prompting 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. To comprehensively compare multiple prompting templates, we divide these templates into base and disambiguation prompting templates.

Model 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 two categories: **(a)** Open-source: we selected 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). **(b)** 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 used the instruction-tuned version, i.e., the chat model, for zero-shot, CoT and few-shot inference. All our experiments were run using 1 $\times$ NVIDIA V100 32G, for different LLM variants. We used vLLM (Kwon et al., 2023) to save inference time. We keep the parameters consistent with those used in previous study (Qian et al., 2024). 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. In addition, considering that GPT is also one of the most popular closed-source models, we further evaluated the performance of GPT-4o-mini. The results are presented in Appendix C.

4 Test set Construction

In this section, we aim to construct a multi-domain ambiguous word vocabulary to annotate the test set. Our annotation processing consists of three steps:

Algorithm 1 Ambiguous Vocabulary Construction

```
Input: Bilingual Vocabulary for multiple domains: V_1, V_2, V_3, V_4, V_5
Output: Ambiguous Vocabulary for each domain: AV_1, AV_2, AV_3, AV_4, AV_5
for each domain i ∈ {1, 2, 3, 4, 5} do
    Initialize AV_i ← ∅
    for each (s, t) ∈ V_i do
        for each j ∈ {1, 2, 3, 4, 5}, j ≠ i do
            if ∃(s, t') ∈ V_j and t' ≠ t then
                Add (s, t) and (s, t') to AV_i
            end if
        end for
    end for
end for
Return: AV_1, AV_2, AV_3, AV_4, AV_5
```

Step 1: Bilingual Vocabulary Construction. In the first step, we utilize *Awesome-Align*[3] (Dou and Neubig, 2021) to perform word alignment on training corpora from multiple domains and obtain a bilingual lexicon. Then, we deduplicate and merge the bilingual vocabulary for each domain based on the source language. Through the above operations, we can obtain ambiguous words within each domain (i.e., *polysemy phenomenon*).

Step 2: Ambiguous Vocabulary Construction. In the second step, we further construct a cross-domain ambiguity vocabulary based on the bilingual vocabulary from the step 1. Algorithm 1 identifies ambiguous vocabulary in multiple domains by comparing bilingual word pairs across domains. For each domain, it initializes an empty ambiguous vocabulary set. Then, for each word pair in the domain, it checks if the source word appears in another domain with a different target translation. If so, both translations are added to the ambiguous vocabulary set. Finally, the algorithm returns the ambiguous vocabulary for each domain. The final vocabulary size is shown in Table 2.

Step 3: Ambiguous Word Annotation. In the third step, we manually annotate the sentences in the test set of each domain based on the ambiguous vocabulary obtained in the second step. The number of one-to-many source language words in the test set is shown in Table 2.

5 Disambiguation Prompting Templates

This section presents multiple prompting strategies combining the domain knowledge (DK).

[3]https://github.com/neulab/awesome-align

English-Chinese					
Train set	Edu	Laws	News	Sci	Sub
	30K	54K	33K	50K	21K
Test set	Edu	Laws	News	Sci	Sub
	492	686	720	471	516
German-English					
Train set	IT	Kor	Laws	Med	Sub
	19K	9K	33K	45K	24K
Test set	IT	Kor	Laws	Med	Sub
	114	67	212	145	109

Table 2: The size of the ambiguous word list in the training and test sets. Edu represents for the Education domain, Sci represents for the Science domain, and Sub represents for the subtitles domain.

5.1 Zero-shot Prompting +DK

In this work, our prompt includes 1) instructions to perform the task such as "*Please translate the following sentence into <target language>*" (i.e., Template 1, T1), and 2) domain-sensitive information such as domain tag. Since Jiao et al. (2023b) have shown that their prompt template can achieve state of-the-art performance using ChatGPT, we mainly followed their template to create our prompt instruction as shown in Figure 2. We design the Zero-shot Prompting +DK with sentence-level and word-level strategies:

Sentence-level. We constructed prompt **Template 5** "*Please translate the following sentence into <target language> according to the <domain tag> domain*". This template mainly utilize the domain information from the sentence-level domain tag.

Word-level. Template 6 "*Please translate the following sentence into <target language> according to the <domain tag> domain of each word*". This template further utilize the domain information of each word. We aim to evaluate whether fine-grained domain information can disambiguate and improve the capability of LLMs' understanding.

5.2 COT Prompting + DK

We also tested 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 4:

Template 7: In this prompt, we give domain tag in step 2. This template further utilize domain information in reasoning ;

Template 8: In this prompt, we ask LLMs to automatically discriminate which domain the source sentence comes from in step 1.

5.3 Few-shot Prompting + DK

We utilize the BM25 (Agrawal et al., 2023) to retrieves the most similar example to the test source from the training datastore. We use 5-shot examples for translation in our main results, this prompt is **Template 3**. To further integrate domain information, we add domain tags to each example, enhancing the LLM's ability to perceive the domain as **Template 9**.

5.4 Reflection Prompting + DK

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 2 **Template 10**.

6 Evaluation Metrics

In this section, we focus on two evaluation metrics: overall translation performance and disambiguation accuracy.

6.1 Metrics of Translation Quality

To better evaluate translation performance, 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[4] scores, the scope is 0-1, for convenience, we multiply the comet score by 100 in our experimental results. In particular, we use the paired bootstrap resampling methods (Koehn, 2004) for the statistical significance test. The specific results are shown in Tables 3 and 4.

Comparison of Base Prompting. Table 3 presents the base prompting results of templates T1-T4 for different LLMs, including LLaMA-3-8B, Qwen-2.5-7B, ALMA-7B, Mistral-7B, and Gemma-2-9B

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

		Education	Laws	News	Science	Subtitles	AVG
ALMA-7B	T1	27.86 / 86.87	23.35 / 88.95	28.57 / 84.02	25.39 / 84.39	21.05 / 78.56	25.24 / 84.56
	T2	30.45 / 87.14	41.43 / 89.24	28.13 / 83.99	25.79 / 84.36	20.71 / 78.57	29.30 / 84.66
	T3	29.64 / 86.86	43.41 / 89.54	27.22 / 83.65	26.00 / 84.51	20.94 / 78.67	29.44 / 84.65
	T4	27.86 / 86.88	24.91 / 89.07	28.26 / 83.93	25.82 / 84.50	21.24 / 78.64	25.62 / 84.60
LLaMA-3-8B	T1	22.97 / 77.40	22.88 / 71.30	16.03 / 72.31	15.81 / 74.04	13.94 / 69.33	18.33 / 72.88
	T2	22.70 / 79.50	31.21 / 73.31	21.32 / 74.89	19.87 / 76.00	16.74 / 71.16	22.37 / 74.97
	T3	28.20 / 86.67	43.27 / 87.67	23.92 / 82.95	22.59 / 84.01	18.94 / 77.69	27.38 / 83.80
	T4	20.37 / 78.76	26.53 / 73.23	17.77 / 76.09	17.67 / 76.69	16.31 / 72.19	19.73 / 75.39
Mistral-7B	T1	14.86 / 77.96	26.01 / 79.96	16.22 / 77.40	15.68 / 78.56	14.63 / 72.67	17.48 / 77.31
	T2	19.04 / 81.53	24.10 / 79.76	15.71 / 77.90	15.09 / 80.34	14.11 / 74.08	17.61 / 78.72
	T3	18.22 / 82.54	26.12 / 82.88	17.01 / 79.63	16.21 / 80.97	15.00 / 75.30	18.51 / 80.26
	T4	10.99 / 74.26	7.38 / 66.07	6.27 / 67.15	7.42 / 70.10	7.81 / 67.35	7.97 / 68.99
Gemma-2-9B	T1	15.62 / 77.05	20.03 / 81.87	15.96 / 78.28	17.66 / 78.54	15.89 / 72.80	17.03 / 77.71
	T2	16.32 / 79.09	20.36 / 83.23	16.56 / 79.51	18.16 / 80.83	10.26 / 65.49	16.33 / 77.63
	T3	18.12 / 81.08	20.66 / 83.35	16.78 / 79.80	18.99 / 82.79	13.69 / 69.35	17.65 / 79.27
	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	21.02 / 78.82	32.76 / 85.04
	T2	**34.02 / 88.06**	**51.19 / 89.60**	**30.51 / 84.91**	**28.82 / 85.91**	**23.14 / 79.26**	**33.54 / 85.55**
	T3	34.17 / 88.17	50.48 / 89.22	29.91 / 84.66	28.33 / 85.64	21.98 / 78.89	32.97 / 85.32
	T4	26.75 / 86.06	47.77 / 87.76	26.16 / 82.71	25.90 / 84.03	19.55 / 77.74	29.23 / 83.66

Table 3: BLEU and COMET scores on the English->Chinese translation task (T1-T4) with different open-source LLMs. The best results are highlighted in bold.

in the English-to-Chinese translation task. Based on the results, we summarize the following key findings. **(1)** Qwen-2.5-7B achieves the highest average BLEU and COMET scores across five domains, demonstrating its superior cross-domain translation capabilities. This highlights its robustness and effectiveness in handling diverse domains, making it the primary model for further experiments. **(2)** Few-shot prompting (T3) performs best on LLaMA-3-8B and ALMA-7B, indicating that providing more in-domain examples significantly enhances domain-specific translation quality. This suggests that example-driven prompting is particularly beneficial for improving multi-domain adaptation. **(3)** Chain-of-Thought (COT) prompting (T2) achieves better results on larger LLMs, such as Qwen-2.5-7B and LLaMA-3-8B, suggesting that reasoning-based prompting more effectively activates domain knowledge, leading to improved translation quality (Jiang et al., 2024). **(4)** While Qwen-2.5-7B exhibits the most consistent cross-domain performance, other models show varying strengths under different prompting strategies. For example, ALMA-7B and LLaMA-3-8B benefit more from Few-shot prompting, whereas Mistral-7B and Gemma-2-9B maintain relatively stable but lower performance across tasks. **(5)** Different LLMs respond differently to prompting strategies, emphasizing the importance of selecting appropriate methods based on model characteristics. Larger models tend to benefit more from COT prompting, while smaller or less generalized models achieve better performance with Few-shot prompting. In summary, Qwen-2.5-7B emerges as the best-performing model for cross-domain translation, and T3 (Few-shot) and T2 (COT) prove to be effective prompting strategies depending on the model's characteristics.

Comparison of Disambiguation Prompting. As shown in the top part of Table 4, for the English->Chinese translation direction, we summarize our findings as follows: (1) Domain knowledge (DK) generally improves translation quality, with the most significant benefits observed in the Chain-of-Thought (COT) and Reflection methods. (2) Zero-shot prompting combined with DK shows inconsistent results: T5 outperforms T1 (+0.55 BLEU), but T6 underperforms (-0.28 BLEU), indicating that DK's impact varies based on the training strategy. (3) Few-shot prompting with DK provides minor improvements, mainly benefiting low-resource domains like Subtitles (+0.47 BLEU), but shows limited gains or even slight degradation in high-resource domains. (4) Reflection with DK achieves the most substantial improvement (+3.77 BLEU on average), particularly in high-resource domains like Education (+6.05 BLEU), Laws (+2.84 BLEU), and News (+4.08 BLEU), suggesting that this approach effectively integrates

English->Chinese							
		Education	**Laws**	**News**	**Science**	**Subtitles**	**AVG**
Zero-shot w/o DK	T1	33.14 / 88.10	50.82 / 88.94	30.04 / 84.51	28.76 / 84.82	21.02 / 78.82	32.76 / 85.04
	T5	33.46 / 88.21	51.39 / 89.20	30.36 / 84.92	28.78 / 86.13	22.53 / 79.35	**33.30/ 85.56**
	T5-T1	+ 0.32 / + 0.11	+ 0.57 / + 0.26	+ 0.32 / + 0.41	+ 0.02 / + 1.31	+ 1.51 / + 0.53	+ 0.55 / + 0.52
	T6	32.64 / 87.84	50.10 / 88.29	30.10 / 84.25	27.99 / 85.50	21.55 / 77.64	32.48 / 84.70
	T6-T1	- 0.50 / - 0.26	- 0.72 / - 0.65	+ 0.06 / - 0.26	- 0.77 / + 0.68	+ 0.53 / - 1.18	- 0.28 / - 0.33
COT w/o DK	T2	34.02 / 88.06	51.19 / 89.60	30.51 / 84.91	28.82 / 85.91	23.14 / 79.26	33.54 / 85.55
	T7	34.50 / 88.09	52.09 / 90.15	31.00 / 85.15	28.97 / 86.05	23.23 / 79.26	**33.96 / 85.74**
	T7-T2	+ 0.48 / + 0.03	+ 0.9 / + 0.55	+ 0.49 / + 0.24	+ 0.15 / + 0.14	+ 0.09 / + 0.00	+ 0.42 / + 0.19
	T8	33.56 / 88.22	50.39 / 88.79	30.15 / 84.88	28.95 / 86.05	22.52 / 79.30	33.11 / 85.45
	T8-T2	- 0.46 /+ 0.16	- 0.8 / -0.81	-0.36 / -0.03	+ 0.13 / + 0.14	-0.62 / + 0.04	- 0.42 / - 0.10
Few-shot w/o DK	T3	34.17 / 88.17	50.48 / 89.22	29.91 / 84.66	28.33 / 85.64	21.98 / 78.89	**32.97 / 85.32**
	T9	33.63 / 88.03	50.46 / 89.49	29.76 / 84.68	27.94 / 85.89	22.45 / 79.04	32.85 / 84.43
	T9-T3	- 0.54 / - 0.14	- 0.02 /+ 0.27	-0.15 / + 0.02	- 0.39 / + 0.25	+ 0.47 / + 0.15	- 0.13 / - 0.98
Reflection w/o DK	T4	26.75 / 86.06	47.77 / 87.76	26.16 / 82.71	25.90 / 84.03	19.55 / 77.74	29.23 / 83.66
	T10	32.80 / 87.83	50.61 / 89.16	30.24 / 84.57	28.60 / 85.68	22.75 / 79.11	**33.00 / 85.27**
	T10-T4	+ 6.05 / + 1.77	+ 2.84 / + 1.40	+ 4.08 / + 1.86	+ 2.70 / + 1.65	+ 3.20 / + 1.37	+ 3.77 / + 1.61

Chinese->English							
		Education	**Laws**	**News**	**Science**	**Subtitles**	**AVG**
Zero-shot w/o DK	T1	22.19 / 83.05	36.03 / 83.48	17.63 / 80.31	16.52 / 81.36	15.34 / 76.46	21.54 / 80.93
	T5	26.61 / 84.02	34.00 / 83.37	18.35 / 80.94	17.68 / 81.86	15.55 / 76.77	**22.44 / 81.39**
	T5-T1	+ 4.42 / + 0.97	+ - 2.03 / - 0.11	+ 0.72 / + 0.63	+ 1.16 / + 0.50	+ 0.21 / + 0.31	+ 0.90 / + 0.46
	T6	25.05 / 83.44	33.42 / 82.67	16.20 / 78.38	16.88 / 80.39	14.26 / 72.52	21.16 / 79.48
	T6-T1	+ 2.86 / + 0.39	- 2.61 / - 0.81	- 1.43 / - 1.93	+ 0.36 / - 0.97	- 1.08 / - 3.94	- 0.38 / - 1.45
COT w/o DK	T2	26.65 / 84.17	33.49 / 83.44	17.82 / 80.6	18.12 / 81.82	15.71 / 76.75	22.36 / 81.36
	T7	26.63 / 84.91	34.46 / 83.72	18.08 / 80.29	18.82 / 81.86	15.83 / 77.97	**22.76 / 81.75**
	T7-T2	- 0.02 / + 0.74	+ 0.97 / + 0.28	+ 0.26/ - 0.31	+ 0.70 / + 0.04	+ 0.12 / + 1.22	+ 0.41 / + 0.39
	T8	26.38 / 84.23	33.88 / 83.54	17.89 / 80.72	18.18 / 81.90	15.25 / 76.73	22.32 / 81.42
	T8-T2	- 0.27 / + 0.06	+ 0.39 / + 0.10	+ 0.07 / + 0.12	+ 0.06 / + 0.08	- 0.46 / - 0.02	- 0.04 / + 0.07
Few-shot w/o DK	T3	26.36 / 84.05	32.08 / 83.44	18.67 / 80.69	18.26 / 81.8	14.53 / 76.72	21.98 / 81.34
	T9	27.11 / 83.96	32.11 / 83.56	18.68 / 80.8	18.33 / 81.91	14.97 / 76.96	**22.24 / 81.44**
	T9-T3	+ 0.75 / - 0.09	+ 0.03 / + 0.12	+ 0.01 / + 0.11	+ 0.07 / + 0.11	+ 0.44 / + 0.24	+ 0.26 / + 0.10
Reflection w/o DK	T4	15.55 / 78.96	28.33 / 80.06	16.07 / 79.35	15.15 / 79.86	13.94 / 74.81	17.81 / 78.61
	T10	24.33 / 83.69	34.25 / 83.77	17.72 / 80.68	16.84 / 81.87	15.85 / 76.73	**21.80 / 81.35**
	T10-T4	+ 8.78 / + 4.73	+ 5.92 / + 3.71	+ 1.65 / + 1.33	+ 1.69 / + 2.01	+ 1.91 / + 1.92	+ 3.99 / + 2.74

Table 4: BLEU and COMET scores on the English-to-Chinese translation task for T1-T10 with Qwen-2.5-7B. We bold the best performance results. Orange text stands for the disambiguation prompting templates.

domain-specific knowledge. Overall, our findings indicate that incorporating DK is beneficial, especially for COT and Reflection methods, making them preferable choices for domain-aware translation tasks.

For the Chinese-to-English translation direction, Incorporating domain knowledge (DK) improves translation performance, especially in Reflection (+3.99 BLEU) and COT (+0.41 BLEU). In zero-shot prompting, DK has mixed effects, with T5 outperforming T1 (+0.90 BLEU) but T6 performing worse (-0.38 BLEU). Few-shot prompting sees minor gains (+0.26 BLEU), while Reflection benefits the most, notably in Education (+8.78 BLEU) and Laws (+5.92 BLEU). Overall, DK is most effective in Reflection and COT, making them preferred choices for domain-specific translation. We further present the results for the German-to-English translation direction, as detailed in Appendix A.

6.2 Metrics of Disambiguate

To evaluate the disambiguation capability of DK, we design two evaluation metrics: Alignment Accuracy and GPT-4o Evaluator. In Table 3, we have shown a strong performance of our approach. To give a better understanding of our DK, we conduct several detailed analyses and discussions in this section on the English->Chinese translation task.

Alignment Accuracy. We count the total number of ambiguous words n in the test set based on the ambiguous word dictionary, and count the num-

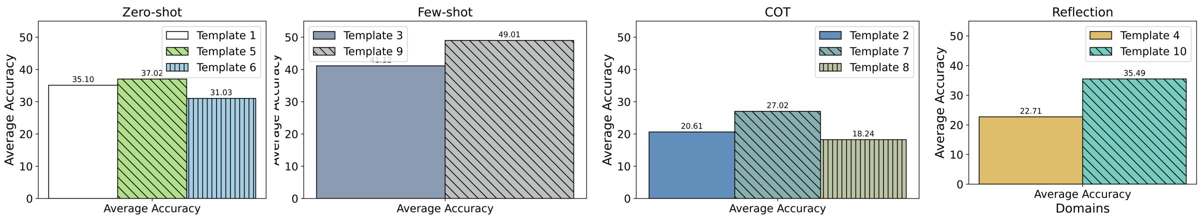

Figure 3: Comparison of Disambiguation Accuracy of Different Templates with GPT-4o-mini Evaluator

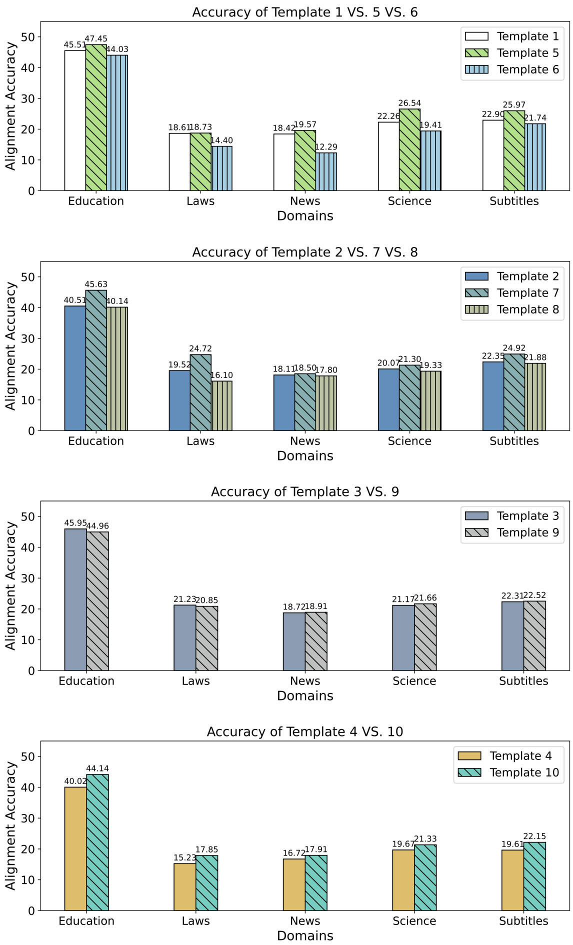

Figure 4: Comparison of Disambiguation Accuracy of Different Templates

ber of correctly translated ambiguous words m to represent alignment accuracy: $\frac{m}{n}$. As shown in Figure 4, by comparing with different base prompting templates, DK templates achieve the performance improvements of alignment accuracy, indicating that DK achieve correct translations of ambiguous words by enhancing domain-specific information, thereby eliminating ambiguity and further improving translation performance.

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 large language models 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 3. DK has improved disambiguation accuracy across the four strategies. The consistency with Figure 4 further proves the effectiveness of disambiguation prompting. We provide specific examples in Appendix B to analyze and illustrate the effectiveness in disambiguation capability.

7 Conclusion

In this paper, we construct a test set with multi-domain ambiguous word annotation, introduce a novel domain disambiguation metric, and investigate various prompting strategies, particularly disambiguation prompting, to evaluate their effectiveness in resolving ambiguities across different domains. Our results demonstrate that integrating domain knowledge and examples significantly improves LLMs' disambiguation accuracy, outperforming traditional prompting methods. We also find that sentence-level domain information and domain-specific examples play a crucial role in MDT. Furthermore, while standard metrics like BLEU and COMET may not directly capture improvements in disambiguation accuracy, they remain valuable for assessing overall translation performance. For future work, we will explore how fine-tuning LLMs with domain-specific data could further enhance their disambiguation and translation quality.

Limitations

Our method was initially validated on models with around 7 billion parameters and has not been tested on larger open-source models. Larger models may potentially yield better performance. Additionally, regarding the evaluator, considering the high economic cost of using GPT-4, we have chosen to use the more cost-effective GPT-4o-mini in this work. We plan to address these limitations and expand on them in future work.

References

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

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.

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*.

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, Ruizhe Chen, Yan Zhang, Zijie Meng, and Zuozhu Liu. 2024. Ladder: A model-agnostic framework boosting llm-based machine translation to the next level. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15377–15393.

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.

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

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, and Zhaopeng Tu. 2023b. Is chatgpt a good translator? a preliminary study. *arXiv 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.

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. 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, Yuanmeng Chen, Yufeng Chen, and Jinan Xu. 2023. Exploring domain-shared and domain-specific knowledge in multi-domain neural machine translation. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 99–110, Macau SAR, China. Asia-Pacific Association for Machine Translation.

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.

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.

Wenbo Pan, Qiguang Chen, Xiao Xu, Wanxiang Che, and Libo Qin. 2023. A preliminary evaluation of chatgpt for zero-shot dialogue understanding. *arXiv preprint arXiv:2304.04256*.

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.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S Yu. 2024. Large language models meet nlp: A survey. *arXiv preprint arXiv:2405.12819*.

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.

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*.

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*.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2024. Delta: An online document-level translation agent based on multi-level memory. *arXiv preprint arXiv:2410.08143*.

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. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

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.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

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*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A German-English Translation Task

To further show the advantages of our method, we compare the results of DK with Qwen-2.5-7B on the German->English translation task. Table 5 presents detailed comparisons, it is worth noting that in the German-to-English translation task, compared to the basic prompts (T1-T4), all five disambiguation prompts (T5-T10), except for T8, showed significant improvements. For example, T5 achieved increases of +4.75 and +4.54 in BLEU and COMET scores, respectively, these results of German-to-English translation tasks further validate the robustness and versatility of our approach DK. Regarding the performance decline of T8. In addition, we observed performance differences for T5 when the target language is Chinese versus German. Specifically, compared to T1, T5 shows a

German->English						
	Education	**Laws**	**News**	**Science**	**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	**29.20 / 83.70**
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 5: BLEU and COMET scores on the German->English translation task for T1-T10 with Qwen-2.5-7B. We bold the best performance results. Orange text stands for the disambiguation prompting

performance decline in Chinese but an improvement in English. We hypothesize that this is partly due to the larger volume of training data available for English in large language models. From a linguistic perspective, English inherently has word segmentation, making it easier to enhance domain-specific information when considering word-level features.

B Case Study

Within the example in Table 6 shows that five translation cases selected from the test datasets in the different domains. We can see that benefiting from the disambiguation prompting mechanism, our approach can generate the correct translation, further showing that our method can effectively learn multiple new domain knowledge.

C Results of GPT-4o-mini

As shown in Table 7, compared to the base prompting, DK has improved both BLEU and COMET scores across different strategies, further demonstrating the stability of our method on GPT-4o-mini.

Domain	Education
Source	Narrow the question down to a coherent and manageable set of issues.
Reference	把问题具体到一系列相关、易处理的主题。
T1	将问题聚焦到一个连贯且manageable 的议题集合上。
T5	将问题聚焦到一组连贯且可管理的问题上。
T6	将问题缩小到一个连贯且可管理的问题范围。
Domain	**Laws**
Source	Torture or inhuman treatment of any resident shall be prohibited.
Reference	禁止对居民施行酷刑或予以非人道的对待。
T2	任何居民不得遭受酷刑或虐待。
T7	虐待或以残忍的方式对待任何居民是被禁止的。
T8	任何居民不得遭受酷刑或残忍、不人道的待遇。
Domain	**Science**
Source	The research involves atmospheric conditions at few levels above the ground.
Reference	此项研究涉及地面之上若干层次的大气状态。
T3	研究涉及地面以上少数几层的气象条件。
T9	此项研究涉及地面之上若干层次的大气状态。
Domain	**News**
Source	You can get whiplash from reading the economic news coming out of China these days.
Reference	最近这些天，阅读中国的经济新闻可能会扭伤你的脖子。
T4	你可以从最近中国的经济新闻报道中感受到whiplash效应。
T10	你可能会从最近有关中国经济的新闻中感受到whipshock的效果。

Table 6: English->Chinese translation cases. Blue indicates the correct translation, while red indicates an incorrect translation.

English->Chinese						
	Education	**Laws**	**News**	**Science**	**Subtitles**	**AVG**
T1	36.77 / 88.74	50.32 / 90.31	33.00 / 85.51	30.91 / 86.61	24.78 / 80.02	35.16 / 86.24
T5	36.90 / 88.79	50.52 / 90.39	33.19 / 85.60	30.93 / 86.77	25.33 / 80.36	**35.37 / 86.38**
T6	36.58 / 88.60	50.47 / 80.42	32.55 / 85.47	30.25 / 86.39	25.24 / 79.98	35.02 / 84.17
T2	37.22 / 88.15	49.81 / 90.19	32.68 / 85.45	31.08 / 86.24	25.44 / 80.35	35.25 / 86.07
T7	38.96 / 88.61	49.86 / 90.21	32.74 / 85.54	32.88 / 86.42	26.64 / 80.10	**36.22 / 86.18**
T8	34.22 / 80.23	45.36 / 85.63	30.26 / 83.00	31.11 / 85.01	23.22 / 77.96	32.83 / 82.37
T3	35.15 / 88.40	52.63 / 90.54	33.12 / 85.48	30.74 / 86.59	24.14 / 80.26	35.16 / 86.25
T9	36.99 / 89.32	52.62 / 90.69	33.25 / 86.55	31.45 / 87.62	25.13 / 80.66	**35.89 / 86.97**
T4	35.97 / 88.50	50.60 / 90.14	32.74 / 85.43	31.24 / 86.76	24.92 / 80.31	35.09 / 86.23
T10	36.71 / 89.12	50.86 / 90.33	33.16 / 86.19	31.42 / 86.01	24.99 / 80.96	**35.43 / 86.52**

Table 7: BLEU and COMET scores on the English->Chinese translation task for T1-T10 with **GPT-4o-mini**. We bold the best performance results. Orange text stands for the disambiguation prompting