How Well Do Large Reasoning Models Translate? A Comprehensive Evaluation for Multi-Domain Machine Translation

Anonymous ACL submission

Abstract

Large language models (LLMs) have demonstrated strong performance in general-purpose machine translation, but their effectiveness in complex, domain-sensitive translation tasks remains underexplored. Recent advancements in Large Reasoning Models (LRMs), raise the question of whether structured reasoning can enhance translation quality across diverse domains. In this work, we compare the performance of LRMs against traditional LLMs and neural MT models across 15 representative domains and four translation directions. Our evaluation considers various factors, including task difficulty, input length, and terminology density. We use a combination of automatic metrics and an enhanced MQM-based evaluation hierarchy to assess translation quality. Our findings show that LRMs consistently outperform traditional LLMs in semantically complex domains, especially in long-text and high-difficulty translation scenarios. Moreover, domain-adaptive prompting strategies further improve performance by better leveraging the reasoning capabilities of LRMs. These results highlight the potential of structured reasoning in MDMT tasks and provide valuable insights for optimizing translation systems in domain-sensitive contexts.

1 Introduction

Large language models (LLMs) have been widely applied to machine translation (MT) and demonstrate superior performance in general-domain settings (Mao et al., 2024; Peng et al., 2023; Jiao et al., 2023; Hendy et al., 2023; Siu, 2023). However, LLMs face persistent challenges in multi-domain MT (MDMT), including instability in domain disambiguation, poor alignment with domain-specific terminology and style, and limited cross-domain generalization (Zheng et al., 2024; Hu et al., 2024; Man et al., 2025). These challenges often require implicit reasoning, such as inferring domain context, selecting appropriate terminology, and maintaining stylistic coherence across varied domains—capabilities that go beyond standard sequence generation.

Recently, a new class of reasoning-oriented LLMs known as large reasoning models (LRMs) has emerged, including models such as OpenAI-o1 (OpenAI et al., 2024b) and DeepSeek-R1 (DeepSeek-AI et al., 2025a). These models incorporate structured generation mechanisms such as multi-step planning, intermediate verification, and self-refinement, and have achieved strong results in complex tasks such as code generation (Zhang et al., 2024b; Quan et al., 2025; Zhang et al., 2025) and mathematical reasoning (Zhao et al., 2025; Huang et al., 2025). Given these capabilities, a natural question arises: ***Can the reasoning abilities of LRMs be leveraged to address the core challenges of MDMT?*** Unfortunately, this question has received little attention, and no prior work has systematically examined the translation capabilities of LRMs in multi-domain settings.

In this work, we aim to fill this gap through a systematic evaluation of LRMs on multi-domain translation benchmarks. Specifically, we compare LRMs with two model classes: LLMs, and neural machine translation (NMT) models. Our evaluation spans 15 representative domains and four translation directions, including 11 domains in the German⇔English setting and 12 in the Chinese⇔English setting, covering both general-purpose (e.g., conversation, news) and specialized (e.g., law, literature) domains.

To capture the diverse challenges of MDMT, we design our evaluation around five key aspects: structural complexity (Section 4.6), long-range dependency modeling (Section 4.4), terminology alignment (Section 4.3), stylistic consistency (Section 4.5), and domain sensitivity (Section 4.2). We assess these aspects using both combines automatic metrics (BLEU, COMET, CometKiwi) and an LLM-powered scoring protocol based on the

Multidimensional Quality Metrics (MQM) hierarchy, allowing us to analyze how reasoning capacity and domain control interact across diverse translation conditions.

Based on our comprehensive evaluation, we summarize the following key findings: **First**, LRMs achieve higher semantic metrics (COMET, CometKiwi) in open-domain settings (e.g., Conversation, News), but lower BLEU in terminology-intensive domains. **Second**, MQM evaluations on De⇒En show that LRMs perform well in semantic accuracy and fluency, but are weaker in stylistic consistency. In terminology accuracy, LRMs outperform traditional LLMs in Zh⇒En but underperform in De⇒En, suggesting that model suitability may vary across language pairs. **Third**, LRMs demonstrate superior adaptability in document-level translation. Their structured reasoning mechanisms help model longer-range dependencies and cross-sentence coherence, yielding more consistent outputs for long-form texts. **Fourth**, we find that incorporating domain-aware prompts can significantly improve their performance. Prompting strategies that encourage implicit domain inference (rather than static domain label) lead to better semantic and stylistic alignment. **Fifth**, we show that LRMs outperform traditional LLMs on complex translation tasks. Their structured reasoning capabilities allow them to maintain semantic integrity even as translation complexity increases.

Overall, our study not only benchmarks the translation capabilities of LRMs across diverse domains, but also sheds light on their unique strengths and limitations compared to traditional LLMs, offering valuable insights for future research.

2 Related Work

2.1 Domain Adaptation in MDMT

Early MDMT approaches relied on explicit domain labels (Saunders, 2022), such as prepending domain tags to source text (Kobus et al., 2017; Tars and Fishel, 2018). To better capture fine-grained domain signals, some methods introduce word-level domain-aware modeling, separating domain-specific and domain-shared contexts for better adaptation to mixed-domain data (Zeng et al., 2018; Su et al., 2021). To handle missing or uncertain labels, Lu and Zhang (2023) propose a Mixture-of-Domain-Experts architecture that dynamically activates shared and domain-specific experts. Wu et al. (2024) further introduce a domain-reasoning-based expert selection mechanism, improving generalization to unseen domains. Additionally, Zeng et al. (2019) introduced the IDDA framework, which leverages bidirectional knowledge distillation for continual cross-domain adaptation.

2.2 Large Reasoning Models

Enhancing the reasoning capabilities of LLMs has become a critical step toward general intelligence. OpenAI's o1 (OpenAI et al., 2024b) has demonstrated strong performance on complex reasoning tasks by aligning structured thinking via reinforcement learning (RL). Building on this foundation, the "Journey Learning" paradigm (Qin et al., 2024) encourages models to emulate human-like exploration strategies, including trial-and-error and reflection, while subsequent work distills o1-style chains-of-thought (CoT) through supervised fine-tuning (SFT) to improve reasoning stability and depth (Huang et al., 2024b; Wang et al., 2025c; Ma et al., 2025; Yu et al., 2025; Trung et al., 2024). RL-based strategies have since been widely adopted. DeepSeek-R1 (DeepSeek-AI et al., 2025a) applies multi-stage RL training and cold-start Long CoT data to enhance reasoning. Gemini2.0-Flash-Thinking (DeepMind, 2025) extends context windows to support long-range reasoning. More broadly, models like Marco-o1 (Zhao et al., 2024) aim to align reasoning capabilities for general-purpose, open-ended problem-solving.

In the context of MT, reasoning-oriented models remain in their early stages but have demonstrated promising potential (Liu et al., 2025). Recent approaches such as DRT (Wang et al., 2025a), R1-T1 (He et al., 2025), DeepTrans (Wang et al., 2025b), and MT-R1-Zero (Feng et al., 2025) employ Long CoT-based SFT and/or RL to explicitly activate the reasoning capabilities of LLMs for translation tasks. However, these methods have mainly focused on general-domain or single-domain settings, and a systematic evaluation of reasoning capabilities in MDMT remains lacking.

2.3 Evaluation for LLM-based MDMT

LLMs have shown strong performance in general-domain MT, prompting growing interest in their applicability to domain-specific settings (Qian et al., 2024; Huang et al., 2024a; Wang et al., 2023a; Zhu et al., 2024). Recent work has extended evaluation to diverse genres and specialized domains. In literary translation, studies based on paragraph-level and full-novel corpora (Thai et al., 2022; Fonteyne

et al., 2020; Zhang et al., 2024a) report fluent outputs but persistent issues in stylistic fidelity, discourse coherence, and pragmatic nuance. In technical domains such as science, law, and IT, evaluations highlight challenges with terminological accuracy and structural rigor (Eschbach-Dymanus et al., 2024; Kleidermacher and Zou, 2025). Benchmarking efforts (Yan et al., 2024a,b) further reveal limited generalization in low-resource and specialized domains. In structured contexts like e-commerce, targeted fine-tuning improves fluency and semantic similarity (Gao et al., 2024).

Although these studies have provided detailed evaluations of traditional LLMs in domain-specific translation, recent advances in LRMs raise the question of whether such models are better suited to the challenges of domain-sensitive MT. Chen et al. (2025a) have recently evaluated LRMs on general MT tasks; however, their effectiveness in the more challenging MDMT remains underexplored. In this work, we fill this gap with a comprehensive evaluation across diverse domains.

3 Evaluation Setting

Dataset. To systematically evaluate the performance differences between traditional LLMs and LRMs in MDMT, we collect test sets for both German⇔English and Chinese⇔English directions, covering representative domains for each. The selected domains span a wide range of linguistic and contextual characteristics, including high-resource domains such as news; terminology-intensive areas like IT, law, medicine, biomedicine, and e-commerce; low-resource domains such as Koranic texts; context-sensitive areas like culture and commonsense; and stylistically demanding content such as literature. This design ensures high diversity in domain coverage, textual style, and semantic complexity, enabling a comprehensive and multidimensional evaluation of model performance. Appendix A provides further details on the dataset, including statistics and data sources.

Metrics. We evaluate model performance using three automatic metrics—BLEU (Post, 2018), COMET, and CometKiwi (Rei et al., 2022a,b)—to capture both lexical overlap and semantic adequacy. BLEU measures surface-level n-gram similarity, while COMET and CometKiwi offer reference-based and reference-free evaluations that better align with human judgment in complex scenarios. In addition, we apply an LLM-powered human evaluation protocol based on the Multidimensional Quality Metrics (MQM) (Lommel et al., 2014; Klubička et al., 2018), which we adapt to account for domain-specific error types and LLM-specific phenomena. Further details, including the MQM hierarchy, error categories, and prompting setup, are provided in Appendix B.

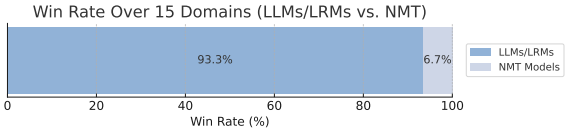

Figure 1: Win rate of LLMs/LRMs vs. NMT across 15 domains. A win is counted if the best LLM or LRM beats the best NMT on three metrics (Table 1). LLMs and LRMs are grouped due to identical results.

Model. We group the evaluated models into two categories. **LLMs.** include GPT-4o (OpenAI et al., 2024a), DeepSeek-V3 (DeepSeek-AI et al., 2025b), and Gemini-2.0-Flash (DeepMind, 2024). **LRMs.** consist of OpenAI-o1 (OpenAI et al., 2024b), OpenAI-o3-mini (OpenAI, 2025), DeepSeek-R1 (DeepSeek-AI et al., 2025a), and Gemini-2.0-Flash-Thinking (DeepMind, 2025). All models are accessed via official APIs provided by OpenAI, DeepSeek, and Gemini (details in Appendix D). **NMT Models.** We use NLLB-200-3.3B (Costa-Jussà et al., 2022) and Google Translate. The former is an open-source MT-specialized model released by Meta, while the latter is a widely-used commercial translation system.

4 Experiments

4.1 Overall Analysis

As shown in Figure 1, LLMs and LRMs collectively outperform NMT models in win rate across 15 domains, underscoring their strong advantage in diverse translation scenarios. We therefore focus the following analysis on comparing LLMs and LRMs.

Across all four translation directions, LRMs exhibit strong performance on semantic metrics (COMET, CometKiwi) in open-domain and user-centric domains such as Conversation, E-commerce, News, and Social. These domains are characterized by natural language usage, diverse expressions, and low terminological rigidity. The structured generation mechanisms of LRMs, optimized for global semantic coherence and human preferences, appear well-suited for such flexible

Model	Biomedical			CommonSense			Conversation			Culture			E-commerce		
	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI
NMT Models															
NLLB-200-3.3B	26.46	81.75	80.9	22.3	78.52	75.14	28.74	82.29	78.37	25.86	76.75	76.44	27.09	81.42	77.82
GoogleTrans	27.02	81.71	80.51	**37.68**	83.63	77.72	29.95	82.9	78.18	27.18	72.9	73.41	29.93	81.1	77.28
Large Language Models															
GPT-4o	32.05	85.89	83.67	32.74	85.25	79.6	**38.68**	87.56	82	38.65	85.24	83.14	**34.17**	86.38	81.78
DeepSeek-v3	32.69	86.09	83.78	33.48	**85.38**	80.01	37.46	87.61	81.99	**40.27**	**85.46**	83.21	33.57	86.32	81.93
Gemini-2.0-Flash	**33.53**	**86.16**	83.6	31.67	84.82	79.16	38.01	87.29	81.87	39.11	85.02	82.94	33.67	86.36	81.67
Large Reasoning Models															
OpenAI-o1	31.3	85.37	**84.97**	28.29	84.97	**81.63**	36.3	**87.87**	**83.49**	34.57	85.06	**84.25**	31.3	**86.39**	**83.05**
DeepSeek-R1	31.48	85.36	83.64	29.27	84.37	80.02	34.05	87.13	81.93	35.88	85.45	83.14	29.23	85.95	81.75
OpenAI-o3-mini	31.4	85.5	83.6	28.65	84.16	80.97	36.33	86.74	81.67	34.05	84.36	**84.25**	32.81	85.92	81.73
Gemini-2.0-Flash-Thinking	32.79	85.58	83.63	32.53	84.53	79.88	37.07	86.99	81.68	36.94	84.63	83.68	33.09	86.17	81.55

Model	IT			Koran			Laws			Literary			Medical		
	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI
NMT Models															
NLLB-200-3.3B	35.02	78.21	76.54	**29.52**	75.91	72.44	34.58	82.6	80.87	11.14	66.97	66.78	40.38	76.99	74.41
GoogleTrans	24.27	75.93	76.51	27.28	**79.28**	75.71	20	73.99	76.71	**20.98**	76.11	74.7	33.35	80.16	78.79
Large Language Models															
GPT-4o	37.23	83.52	78.37	17.65	75.04	80.64	58.99	89.06	84.38	19.24	**78.18**	77.12	41.89	84.23	81.89
DeepSeek-v3	**38.16**	**83.9**	78.86	17.94	74.91	80.33	**59.57**	**89.15**	84.42	18.4	77.94	76.88	41.43	84.15	81.9
Gemini-2.0-Flash	37.93	83.2	78.49	19.7	74.97	79.71	48.5	88.28	84.52	15.85	76.46	76.62	**44.39**	**84.49**	81.76
Large Reasoning Models															
OpenAI-o1	36.4	83.2	**79.75**	16.76	75.03	**82.02**	56.43	88.9	84.4	16.99	77.8	**77.56**	40.44	83.88	**83.29**
DeepSeek-R1	36.66	83.49	78.68	17.05	74.76	80.63	55.75	88.61	84.26	13.11	76.92	76.77	40.23	83.82	81.81
OpenAI-o3-mini	37.13	82.63	78.2	15.52	73.54	80.19	48.11	88.32	**85.13**	17.05	76.92	76.84	39.82	83.6	81.79
Gemini-2.0-Flash-Thinking	37.26	82.94	78.4	19.57	75.03	79.79	42.73	87.66	84.44	17.43	77.22	76.81	43.36	84.14	81.7

Model	Mixed			News			Science			Social			Subtitles		
	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI
NMT Models															
NLLB-200-3.3B	28.07	81.06	77.51	22.81	78.21	73.19	19.91	78.26	73.93	25.94	79.22	76.17	23.21	75.64	72.11
GoogleTrans	31.58	80.91	77.58	28.91	79.19	77.23	25.45	79.36	77.61	30.81	79.57	76.57	24.6	77.34	75.67
Large Language Models															
GPT-4o	**35.38**	86.11	81.98	34.13	85.79	82.86	32.09	87.34	84.84	33.07	84.63	81.47	27.26	81.27	82.09
DeepSeek-v3	34.92	86.19	82.19	34.02	86.08	83.25	**33.54**	**88.28**	85.46	32.96	84.74	81.81	26.68	**81.56**	82.47
Gemini-2.0-Flash	35.25	86.06	81.96	**34.51**	85.94	83.03	33.47	87.69	85.24	**33.24**	84.55	81.49	**27.52**	81.32	82.26
Large Reasoning Models															
OpenAI-o1	32.42	**86.3**	**83.46**	31.67	**86.11**	**84.45**	32.32	87.92	**85.56**	29.95	**84.8**	**82.96**	25.93	81.55	**83.12**
DeepSeek-R1	30.77	85.79	82.07	30.68	85.85	83.24	31.82	88.08	85.47	28.51	84.26	81.59	24.66	81.38	82.53
OpenAI-o3-mini	33.57	85.54	81.78	32.18	85.42	82.73	31.38	87.42	85.24	31.88	84.07	81.28	25.67	80.98	82.23
Gemini-2.0-Flash-Thinking	34.43	85.84	81.83	33.74	85.92	83	31.98	87.47	85.06	32.3	84.16	81.27	27.11	80.84	81.71

Table 1: Translation performance across 15 domains, averaged over four translation directions (En⇔Zh, De⇔En). Metrics include BLEU, COMET, and CometKiwi (KIWI). **Bold** denote the best scores.

translation scenarios.

However, in structurally constrained and terminology-intensive domains like Law, Medical, IT, and Koran, LRMs underperform compared to traditional LLMs, particularly on COMET and BLEU. This performance gap can be attributed to two main factors: first, LRMs tend to prioritize semantic plausibility over strict terminological fidelity, occasionally producing paraphrases or near-synonyms that diverge from reference terms; second, their reasoning process may lead to over-interpretation of domain cues (Chen et al., 2025b), resulting in translations that are semantically valid but lexically mismatched.

Despite these challenges, LRMs maintain strong results on the reference-free CometKiwi metric across all domains, especially OpenAI-o1, which consistently outperforms traditional LLMs. This indicates that LRMs produce translations that are semantically consistent and well-aligned with human expectations, even when deviating from the reference surface form. Notably, in the Zh⇒En Culture domain (Table 15), DeepSeek-V3 achieves better performance than other models, likely due to its Chinese-centric pretraining, which enhances its ability to capture socio-cultural and idiomatic nuance. This underscores the importance of pre-training data alignment in handling linguistically and culturally sensitive content.

Overall, traditional LLMs achieve higher BLEU scores, especially in domains requiring strict formatting and domain-specific terminology. This advantage may stem from supervised fine-tuning on parallel corpora, which promotes reference imitation and boosts surface-level overlap. In contrast, LRMs are often trained with reinforcement

Model	MQM Score (↓)					
	IT	Koran	Medical	Law	Subtitles	Average
GPT-4o	3.14	4.56	3.24	3.74	3.44	3.62
DeepSeek-V3	**2.91**	4.62	3.30	3.89	**3.38**	3.62
Gemini-2.0-Flash	3.09	4.68	**3.15**	**3.55**	**3.38**	**3.57**
OpenAI-o1	3.11	4.74	3.32	3.86	3.49	3.70
DeepSeek-R1	2.92	**4.55**	3.30	3.88	3.40	3.61
OpenAI-o3-mini	3.16	5.22	3.43	4.59	3.62	4.00
Gemini-2.0-Flash-Thinking	3.15	4.84	3.28	3.69	3.41	3.67

Table 2: MQM scores for different models across various domains in the Multi-Domain dataset (De⇒En).

Model	Acc. (%)	Style (%)	Fluency (%)	Term. (%)	Non-trans (%)
GPT-4o	50.78	30.58	7.03	11.38	0.23
DeepSeek-V3	50.72	30.51	6.84	11.83	0.09
Gemini-2.0-Flash	52.86	**29.43**	6.87	**10.68**	0.16
OpenAI-o1	50.95	29.58	7.10	12.22	0.15
DeepSeek-R1	50.50	30.83	6.68	11.91	**0.08**
OpenAI-o3-mini	**49.97**	30.48	7.19	12.24	0.12
Gemini-2.0-Flash-Thinking	52.54	29.64	**6.38**	11.29	0.15

Table 3: Error category percentages for different models, including accuracy (Acc.), style, fluency, terminology (Term.), and non-translation error (Non-trans).

learning objectives favoring semantic alignment, leading to lower BLEU but higher COMET and CometKiwi scores. Importantly, lower BLEU scores in LRMs do not necessarily reflect translation errors; rather, they often indicate semantically faithful but lexically divergent outputs. This reinforces the need for multi-dimensional evaluation. To this end, we present MQM-based fine-grained analysis in the next section.

4.2 MQM Evaluation

To systematically evaluate translation errors across domains, we adopt the MQM hierarchy and employ DeepSeek-V3 as an automatic scoring agent to ensure consistent annotation quality. A structured prompt guides the model to identify up to five major errors per sentence, and scores are computed using severity weights proposed by Freitag et al. (2021). This approach balances scalability and consistency, allowing fine-grained comparison across models on seven key dimensions: Accuracy, Fluency, Style, Terminology, Others, Source Error, and Non-translation Error. In MQM score calculation and error category statistics, Source Error are excluded (Freitag et al., 2021).

As shown in Table 2, traditional LLMs achieve lower overall MQM scores, showing higher consistency and precision. Interestingly, LRMs achieve lower MQM scores in the Koran domain, which may benefit from their explicit CoT reasoning paradigm when handling semantically ambiguous or stylistically atypical text. However, their performance lags behind in terminology-intensive domains such as Medical and Law. Further analysis in Table 3 reveals that LRMs have slightly lower error rates in Accuracy, Fluency, and Non-translation Error, suggesting strong capabilities in semantic similarity and sentence structure. In contrast, they exhibit higher error rates in Style and Terminology—particularly OpenAI-o3-mini and DeepSeek-R1—reflecting weaker handling of domain-specific consistency and stylistic coherence.

These findings highlight the need for future work to further strengthen LRMs' capabilities in deep domain understanding, contextualized terminology selection, and adaptation to distinct domain-specific translation styles, to enhance their robustness in domain-diverse translation scenarios.

Model	Term Acc. (%)	
	Zh⇒En	De⇒En
GPT-4o	52.95	37.67
DeepSeek-V3	51.84	37.92
Gemini-2.0-Flash	52.21	**38.55**
OpenAI-o1	52.92	37.42
DeepSeek-R1	**53.20**	37.83
OpenAI-o3-mini	45.59	37.38
Gemini-2.0-Flash-Thinking	52.67	37.62

Table 4: Terminology accuracy (%) for Zh⇒En and De⇒En translation tasks.

4.3 Terminological Accuracy

In the previous section, we observed that LRMs underperform traditional LLMs in terminology-heavy domains such as law and medicine. This observation prompted a focused investigation into their terminology translation capabilities. To this end, we conducted an evaluation based on the WMT23 Terminology Shared Task (Semenov et al., 2023), which provides bilingual sentence pairs along with corresponding term alignments. For each test example, we input only the source sentence into the model and assess whether the expected target term appears correctly in the generated translation. We report terminology accuracy—the proportion of reference target terms correctly generated in the translation—in both Zh⇒En and De⇒En directions, as shown in Table 4.

In the Zh⇒En setting, LRMs demonstrate superior performance, with DeepSeek-R1 achieving the highest accuracy (53.20%). We attribute this to LRMs' stronger reasoning and contextual comprehension capabilities, which allow them to better resolve semantic ambiguity in Chinese and plan conceptual mappings into English, ensuring the naturalness and accuracy of terminology in the En-

Model	Conversation			E-commerce			News			Social			Literary		
	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI
Document-Level															
GPT-4o	**40.00**	84.71	82.02	35.33	87.35	64.55	36.07	86.60	66.01	33.75	84.36	68.70	17.55	**77.57**	77.14
DeepSeek-V3	39.28	84.51	81.98	36.71	87.26	63.16	37.59	86.50	65.35	36.48	84.80	67.83	16.81	77.48	77.17
Gemini-2.0-Flash	39.98	84.64	81.81	**37.72**	87.34	62.33	38.68	85.71	63.68	**37.76**	84.38	67.14	**18.12**	77.30	76.62
OpenAI-o1	37.05	**84.83**	**82.99**	30.68	**87.40**	**66.50**	32.08	**86.67**	**67.90**	29.02	**84.82**	**69.79**	16.44	**77.57**	**77.76**
DeepSeek-R1	33.66	84.15	81.45	21.62	85.23	63.77	23.49	83.68	66.99	22.22	81.59	68.09	13.12	76.64	76.47
OpenAI-o3-mini	36.58	84.52	82.05	32.13	87.15	65.18	33.60	86.22	65.77	30.93	84.76	68.87	17.26	76.87	76.64
Gemini-2.0-Flash-Thinking	**40.00**	84.59	81.79	37.41	87.35	62.39	**38.82**	86.62	64.19	36.97	84.55	66.47	17.77	77.17	76.58
Sentence-Level															
GPT-4o	**38.68**	87.56	82.00	**34.17**	86.38	81.78	36.21	86.05	82.66	33.07	84.63	81.47	18.53	**82.96**	**57.23**
DeepSeek-V3	37.46	87.61	81.99	33.57	86.32	81.93	35.82	86.30	83.04	32.96	84.74	81.81	19.92	82.88	54.68
Gemini-2.0-Flash	38.01	87.29	81.87	33.67	86.36	81.67	**36.48**	86.21	82.80	**33.24**	84.55	81.49	23.28	82.80	53.83
OpenAI-o1	36.30	**87.87**	**83.49**	31.30	**86.39**	**83.05**	33.17	**86.35**	**84.35**	29.95	**84.80**	**82.96**	16.87	82.42	55.73
DeepSeek-R1	34.05	87.13	81.93	29.23	85.95	81.75	31.83	86.03	83.01	28.51	84.26	81.59	5.16	80.35	52.16
OpenAI-o3-mini	36.33	86.74	81.67	32.81	85.92	81.73	33.95	85.57	82.43	31.88	84.07	81.28	18.06	82.83	55.16
Gemini-2.0-Flash-Thinking	37.07	86.99	81.68	33.09	86.17	81.55	35.70	86.19	82.81	32.30	84.16	81.27	**23.52**	82.77	53.69

Table 5: Average BLEU, COMET, and CometKiwi scores at document-level and sentence-level across five domains.

glish context. These results suggest that LRMs hold strong potential in terminology translation. In contrast, traditional LLMs outperform LRMs in the De⇒En direction, with Gemini-2.0-Flash achieving the highest accuracy (38.55%). This may be attributed to the high lexical similarity between German and English, which allows for more direct term mapping, favoring models trained to mimic reference terms without requiring deep reasoning.

Therefore, in practical applications, model selection should take into account the linguistic gap between source and target languages: LRMs are better suited for language pairs with significant structural differences (*e.g.*, Zh⇒En), while traditional LLMs may be more effective for structurally similar pairs like De⇒En.

Given these findings, we argue that improving the terminology translation performance of LRMs requires tighter integration between their reasoning structures and terminology modeling capabilities. Future work may explore embedding terminology constraints into intermediate reasoning chains to support explicit term-level planning. These structural and reasoning-aligned improvements are expected to unlock the potential of LRMs in terminology-intensive translation scenarios.

4.4 Document-Level Evaluation

To investigate the impact of textual granularity on model performance, we compare sentence-level and document-level inputs across five domains, as shown in Table 5. Compared to traditional LLMs, LRMs demonstrate stronger adaptability in document-level translation. They consistently achieve higher scores on COMET and CometKiwi, which focus on semantic similarity and contextual coherence. This advantage is particularly evident in long-text scenarios, where modeling inter-sentential dependencies and broader discourse structures is essential. Recent studies have shown that LRMs leverage Long CoT reasoning to enhance translation quality in complex texts, effectively modeling longer-range dependencies and maintaining coherence across extended contexts (Liu et al., 2025). In domains like Conversation and News, which naturally involve multi-turn interactions and cross-sentence semantic dependencies, LRMs are better equipped to integrate information across spans and plan discourse globally, enabling both lexical and semantic alignment in context-sensitive scenarios.

These findings highlight the advantages of LRMs for document-level translation. Their capacity for planning, reasoning, and generating globally coherent outputs makes them better suited for long-form and context-sensitive scenarios. Future research could further explore the adaptive translation capabilities of LRMs across different document types, such as legal contracts, technical manuals, and literary works.

4.5 Effect of Domain-Aware Prompt

To further investigate the underperformance of LRMs in domain-intensive settings such as Law, Medical, and IT, we hypothesize that these models may lack sufficient domain sensitivity during translation reasoning, potentially due to the limitations imposed by the prompts in accurately guiding the model to recognize and adapt to domain-specific characteristics. Therefore, we experiment with three prompting strategies (Table 6): P1, where no domain-specific information is provided; P2, where explicit domain labels are given; and P3, where the model is encouraged to infer the domain

Source: Mitteilung an die Kommission und die Mitgliedstaaten.		Reference: **Communication** to the Commission and the Member States.
P1	Reasoning: .....First, "Mitteilung" is the main noun here. The standard translation for that would be "notification" or "communication." Depending on the context, maybe "message," but in formal settings like official documents,"notification" seems more likely..... ***Lacks domain context—defaults to literal mapping***	Prediction: **Notification** to the Commission and the Member States.
P2	Reasoning: .....translated into English in a legal domain style..... "Mitteilung" translates to "communication" or "notification." In legal contexts, "notification" is often more precise.....Now, considering legal terminology, "Mitteilung" here is likely an official notification. So "Notification to the Commission and the Member States" sounds appropriate. I should check if "Member States" is capitalized in legal documents.....That maintains the formal tone and correct terminology..... ***Legal domain injected, but still misaligned—insufficient granularity***	Prediction: **Notification** to the Commission and the Member States.
P3	Reasoning: .....First, "Mitteilung" translates to "communication" or "notification." In an official context, like a document or formal correspondence, "Communication" might be more appropriate.....I should check if there's a standard phrasing used in official EU documents.....Wait, the user mentioned maintaining the domain style of the source language text. The original is formal and bureaucratic, so the translation should mirror that. No need for colloquial terms. Let me verify each part again.....Yes, that's correct. So the final translation..... ***LRM infers EU document style and selects idiomatic legal phrasing***	Prediction: **Communication** to the Commission and the Member States.

Figure 2: Case study of DeepSeek-R1 under three prompting strategies (P1, P2, P3), illustrating its reasoning process and final translation in each setting.

and adjust its translation style accordingly.

Results Analysis. Using DeepSeek-R1 as a representative model, we evaluate the three prompting strategies across IT, Law, and Medical domains. As shown in Table 7, P3 achieves the highest COMET and CometKiwi scores across all domains, indicating its effectiveness in producing semantically faithful and stylistically coherent translations. Both P2 and P3 significantly outperform P1 on semantic metrics, indicating that incorporating domain awareness—either explicitly or implicitly—substantially improves contextual understanding and semantic coherence. While P1 and P3 yield similar BLEU scores in the IT domain, P3 substantially outperforms P1 in Law and Medical, especially in CometKiwi, suggesting that domain-aware reasoning contributes to improved contextual understanding. Interestingly, P2 achieves the highest BLEU scores in the Law domain, likely likely because its explicit domain cues align closely with the rigid structure and formulaic phrasing typical of legal texts. However, P2 lags behind P3 on COMET and CometKiwi, suggesting that fixed labels may limit contextual flexibility. In contrast, P3 allows the model to infer domain style from context, resulting in more fluent and semantically coherent translations.

Case Analysis. To further illustrate the practical impact of prompt design, we present a qualitative example from the legal domain. Given the source text, only P3 yields the reference-consistent translation "Communication" which aligns with standard EU legal terminology. P1 produces "Notification" reflecting a surface-level lexical mapping that lacks sensitivity to domain-specific conventions. P2 also outputs "Notification" despite access to explicit domain labels, suggesting that static domain conditioning alone is insufficient to guide fine-grained lexical selection. In contrast, P3 demonstrates contextualized reasoning: it identifies the formal and bureaucratic nature of the source text and adjusts the output accordingly.

ID	Prompt
P1	Translate the following `{src_lang}` sentence into `{tgt_lang}`.
P2	Given the `{src_lang}` sentence from a `{domain}`-related text, translate it into `{tgt_lang}` in a `{domain}` domain style.
P3	Translate the following `{src_lang}` text into `{tgt_lang}` while maintaining the domain style of the source text.

Table 6: Prompts used in domain-aware translation.

Prompt	IT			Law			Medical		
	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI
P1	**36.66**	83.49	78.68	39.16	84.97	82.35	40.23	83.82	81.81
P2	34.20	82.94	79.33	**41.46**	85.01	83.46	38.92	83.38	82.57
P3	36.55	**83.57**	**79.82**	40.57	**85.32**	**83.54**	**41.14**	**84.06**	**82.97**

Table 7: Translation performance across IT, Law, and Medical domains under different prompt strategies.

These findings highlight the superiority of domain-adaptive prompting over static label specification in MDMT. Future work may focus on injecting domain inference into the reasoning chain, enabling LRMs to autonomously determine the domain and adjust their translation accordingly.

4.6 Impact of Translation Difficulty

In recent years, researchers have proposed various reasoning-enhanced strategies to improve the performance of LLMs on complex tasks (Wei et al., 2022; Yao et al., 2023; Zhou et al., 2022). These approaches focus on guiding models through structured intermediate reasoning to better capture semantic and logical complexity. When extended to the field of MT, we hypothesize that incorporating mechanisms such as problem decomposition, reasoning path exploration, and stepwise inference can

Model	Level 1			Level 2			Level 3			Level 4			Level 5		
	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI
GPT-4o	**41.96**	**85.13**	79.80	30.08	84.88	82.15	26.49	82.75	80.91	22.61	83.11	75.48	24.94	83.59	67.11
DeepSeek-V3	40.81	85.09	79.90	27.90	84.86	82.20	25.43	82.54	80.97	22.57	82.95	75.30	25.06	83.30	67.64
Gemini-2.0-Flash	41.15	84.39	79.46	29.22	84.93	81.96	27.53	82.80	80.68	24.53	82.72	74.57	28.93	83.91	66.28
OpenAI-o1	40.34	84.59	**81.10**	29.67	84.88	**83.28**	25.76	**82.94**	**81.94**	21.99	**83.31**	**76.92**	22.78	**84.02**	**69.62**
DeepSeek-R1	41.43	84.33	79.86	27.12	84.25	81.75	24.32	82.15	80.44	18.66	82.17	75.27	19.53	82.43	68.78
OpenAI-o3-mini	40.74	84.12	79.44	26.49	84.56	82.14	24.93	82.36	80.73	21.09	82.67	75.37	22.70	83.27	67.63
Gemini-2.0-Flash-Thinking	39.87	83.84	79.36	**30.78**	**84.98**	82.21	**28.15**	82.68	80.62	**24.54**	83.02	74.71	**28.97**	83.93	66.32

Table 8: Performance across different levels of complexity, comparing LRMs and tranditional LLMs.

enhance a model's generalization ability in tasks with high semantic load. The LRMs exemplify this reasoning-oriented design philosophy. These models are trained with structured cognitive capabilities, which allows them to operate more robustly in linguistically complex translation scenarios.

To systematically assess the impact of translation difficulty on model performance, we use DeepSeek-V3 to rate the complexity of source sentences drawn from three datasets: Multi-Domain (De⇒En), WMT22 (De⇔En and Zh⇔En), and Guofeng WebNovel (Zh⇒En). These sentences are classified into five difficulty levels, ranging from Level 1, representing the easiest, to Level 5, representing the most difficult. The prompt used for difficulty assessment is provided in Figure 4. As shown in Table 8, traditional LLMs perform well on simpler inputs, especially at Level 1 where they achieve high BLEU and COMET scores. However, their performance declines significantly as difficulty increases. Starting at Level 2, they are consistently outperformed by LRMs. In particular, OpenAI-o1 and Gemini-2.0-Flash-Thinking demonstrate stronger robustness and semantic consistency under higher difficulty conditions, maintaining clear advantages on metrics such as COMET and CometKiwi. This trend supports our hypothesis that LRMs are better equipped to model and generalize in semantically dense and contextually demanding translation tasks.

Future research should further investigate how LRMs adapt to varying levels of translation difficulty. For example, LRMs should generate accurate translations efficiently with minimal reasoning for simple inputs to avoid unnecessary overthinking (Chen et al., 2025b), whereas for more complex sentences, they should engage in deeper analysis and structured reasoning to ensure semantic accuracy and completeness.

Model	BLEU	COMET	KIWI	Avg. Len.
GPT-4o	37.40	86.74	84.61	31.76
DeepSeek-V3	**43.34**	**87.76**	85.04	**29.63**
Gemini-2.0-Flash	40.30	87.05	84.80	32.71
OpenAI-o1	36.17	87.11	**85.09**	402.64
DeepSeek-R1	41.12	87.50	85.03	485.05
OpenAI-o3-mini	34.60	86.69	84.98	424.31
Gemini-2.0-Flash-Thinking	39.17	86.63	84.37	1043.59

Table 9: Average scores and length for various models.

5 Token Consumption Comparison

We compare the average token usage of traditional LLMs and LRMs on the UM-Corpus (En⇒Zh) multi-domain test sets to assess translation efficiency. As shown in Table 9, LRMs generate significantly longer outputs than traditional LLMs. While they achieve strong CometKiwi scores, their lower performance on COMET and BLEU suggests that LRMs may "overthink" during generation (Chen et al., 2025b). In MDMT, future work could explore adapting reasoning depth to translation complexity to reduce overthinking.

6 Conclusion

This paper presents a comprehensive evaluation of LRMs in MDMT. Through systematic comparison with traditional LLMs across 15 domains and four translation directions, we observe that LRMs achieve superior performance in semantically challenging and long-form translation tasks, as reflected in meaning-oriented metrics such as COMET and CometKiwi. However, their performance declines in terminology-intensive and stylistically constrained domains, where traditional LLMs exhibit stronger lexical precision and structural fidelity. Our prompt-based analysis further demonstrates the advantage of domain-adaptive prompting in improving contextual alignment. These findings suggest that future work should focus on integrating structured reasoning with domain-specific constraints to enhance the reliability and controllability of LRMs in real-world translation scenarios.

Limitations

Despite providing a comprehensive evaluation of LRMs in MDMT, our study has several limitations. First, the evaluation is restricted to high-resource language pairs, which may limit the generalizability of our findings to low-resource or typologically distant languages. Second, our automatic MQM annotations rely on a single model as the scorer, which ensures consistency but may introduce bias aligned with that model's own translation behavior.

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.

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy. Association for Computational Linguistics.

Andong Chen, Yuchen Song, Wenxin Zhu, Kehai Chen, Muyun Yang, Tiejun Zhao, and Min zhang. 2025a. Evaluating o1-like llms: Unlocking reasoning for translation through comprehensive analysis. *Preprint*, arXiv:2502.11544.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025b. Do not think that much for 2+3=? on the overthinking of o1-like llms. *Preprint*, arXiv:2412.21187.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Google DeepMind. 2024. Introducing gemini 2.0: our new ai model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#ceo-message. Accessed: 2025-04-21.

Google DeepMind. 2025. Gemini 2.0 flash thinking. https://deepmind.google/technologies/gemini/flash-thinking. Accessed: 2025-04-21.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025b. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Johannes Eschbach-Dymanus, Frank Essenberger, Bianka Buschbeck, and Miriam Exel. 2024. Exploring the effectiveness of llm domain adaptation for business it machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 610–622.

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*.

Margot Fonteyne, Arda Tezcan, and Lieve Macken. 2020. Literary machine translation under the magnifying glass: Assessing the quality of an NMT-translated detective novel on document level. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3790–3798, Marseille, France. European Language Resources Association.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Dehong Gao, Kaidi Chen, Ben Chen, Huangyu Dai, Linbo Jin, Wen Jiang, Wei Ning, Shanqing Yu, Qi Xuan, Xiaoyan Cai, and 1 others. 2024. Llms-based machine translation for e-commerce. *Expert Systems with Applications*, 258:125087.

Jie He, Tao Wang, Deyi Xiong, and Qun Liu. 2020. The box is in the pen: Evaluating commonsense reasoning in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3662–3672, Online. Association for Computational Linguistics.

Minggui He, Yilun Liu, Shimin Tao, Yuanchang Luo, Hongyong Zeng, Chang Su, Li Zhang, Hongxia Ma, Daimeng Wei, Weibin Meng, and 1 others. 2025. R1-t1: Fully incentivizing translation capability in llms via reasoning learning. *arXiv preprint arXiv:2502.19735*.

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 F. Wong, and Rui Wang. 2024. 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.

Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, and 1 others. 2025. Mathperturb: Benchmarking llms' math reasoning abilities against hard perturbations. *arXiv preprint arXiv:2502.06453*.

Xu Huang, Zhirui Zhang, Xiang Geng, Yichao Du, Jiajun Chen, and Shujian Huang. 2024a. Lost in the source language: How large language models evaluate the quality of machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3546–3562, Bangkok, Thailand. Association for Computational Linguistics.

Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. 2024b. O1 replication journey – part 2: Surpassing o1-preview through simple distillation, big progress or bitter lesson? *Preprint*, arXiv:2411.16489.

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

Hannah Calzi Kleidermacher and James Zou. 2025. Science across languages: Assessing llm multilingual translation of scientific papers. *Preprint*, arXiv:2502.17882.

Filip Klubička, Antonio Toral, and Víctor M Sánchez-Cartagena. 2018. Quantitative fine-grained human evaluation of machine translation systems: a case study on english to croatian. *Machine Translation*, 32(3):195–215.

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.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Sinuo Liu, Chenyang Lyu, Minghao Wu, Longyue Wang, Weihua Luo, Kaifu Zhang, and Zifu Shang. 2025. New trends for modern machine translation with large reasoning models. *Preprint*, arXiv:2503.10351.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Jinliang Lu and Jiajun Zhang. 2023. Towards unified multi-domain machine translation with mixture of domain experts. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3488–3498.

Zexiong Ma, Chao Peng, Pengfei Gao, Xiangxin Meng, Yanzhen Zou, and Bing Xie. 2025. Sorft: Issue resolving with subtask-oriented reinforced fine-tuning. *Preprint*, arXiv:2502.20127.

Zhibo Man, Yuanmeng Chen, Yujie Zhang, Yufeng Chen, and Jinan Xu. 2025. Dmdteval: An evaluation and analysis of llms on disambiguation in multi-domain translation. *arXiv preprint arXiv:2504.20371*.

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2024. GPTEval: A survey on assessments of ChatGPT and GPT-4. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7844–7866, Torino, Italia. ELRA and ICCL.

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Cristian Grozea, Amy Siu, Madeleine Kittner, and Karin Verspoor. 2018. Findings of the WMT 2018 biomedical translation shared task: Evaluation on Medline test sets. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 324–339, Belgium, Brussels. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. Openai o3-mini. `https://openai.com/index/openai-o3-mini/`. Accessed: 2025-04-21.

OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec

Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, and 243 others. 2024b. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

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.

Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, and Pengfei Liu. 2024. O1 replication journey: A strategic progress report – part 1. *Preprint*, arXiv:2410.18982.

Shanghaoran Quan, Jiaxi Yang, Bowen Yu, Bo Zheng, Dayiheng Liu, An Yang, Xuancheng Ren, Bofei Gao, Yibo Miao, Yunlong Feng, and 1 others. 2025. Codeelo: Benchmarking competition-level code generation of llms with human-comparable elo ratings. *arXiv preprint arXiv:2501.01257*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Danielle Saunders. 2022. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *Preprint*, arXiv:2104.06951.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the WMT 2023 shared task on machine translation with terminologies. In *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore. Association for Computational Linguistics.

Sai Cheong Siu. 2023. Chatgpt and gpt-4 for professional translators: Exploring the potential of large language models in translation. *Available at SSRN 4448091*.

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.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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).

Luong Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. ReFT: Reasoning with reinforced fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7601–7614, Bangkok, Thailand. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2025a. Drt: Deep reasoning translation via long chain-of-thought. *Preprint*, arXiv:2412.17498.

Jiaan Wang, Fandong Meng, and Jie Zhou. 2025b. Deep reasoning translation via reinforcement learning. *Preprint*, arXiv:2504.10187.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023a.

Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023b. Findings of the WMT 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine Translation*, pages 55–67, Singapore. Association for Computational Linguistics.

Ruida Wang, Rui Pan, Yuxin Li, Jipeng Zhang, Yizhen Jia, Shizhe Diao, Renjie Pi, Junjie Hu, and Tong Zhang. 2025c. Ma-lot: Multi-agent lean-based long chain-of-thought reasoning enhances formal theorem proving. *arXiv preprint arXiv:2503.03205*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Junhong Wu, Yuchen Liu, and Chengqing Zong. 2024. F-MALLOC: Feed-forward memory allocation for continual learning in neural machine translation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7180–7192, Mexico City, Mexico. Association for Computational Linguistics.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Jing Li, Xianchao Zhu, and Yue Zhang. 2024a. Benchmarking gpt-4 against human translators: A comprehensive evaluation across languages, domains, and expertise levels. *Preprint*, arXiv:2411.13775.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024b. Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *Preprint*, arXiv:2407.03658.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Bin Yu, Hang Yuan, Yuliang Wei, Bailing Wang, Weizhen Qi, and Kai Chen. 2025. Long-short chain-of-thought mixture supervised fine-tuning eliciting efficient reasoning in large language models. *Preprint*, arXiv:2505.03469.

Jiali Zeng, Yang Liu, Jinsong Su, Yubing Ge, Yaojie Lu, Yongjing Yin, and Jiebo Luo. 2019. Iterative dual domain adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 845–855, Hong Kong, China. Association for Computational Linguistics.

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.

Alexander Zhang, Marcus Dong, Jiaheng Liu, Wei Zhang, Yejie Wang, Jian Yang, Ge Zhang, Tianyu Liu, Zhongyuan Peng, Yingshui Tan, and 1 others. 2025. Codecriticbench: A holistic code critique benchmark for large language models. *arXiv preprint arXiv:2502.16614*.

Ran Zhang, Wei Zhao, and Steffen Eger. 2024a. How good are llms for literary translation, really? literary translation evaluation with humans and llms. *arXiv preprint arXiv:2410.18697*.

Yuxiang Zhang, Shangxi Wu, Yuqi Yang, Jiangming Shu, Jinlin Xiao, Chao Kong, and Jitao Sang. 2024b. o1-coder: an o1 replication for coding. *Preprint*, arXiv:2412.00154.

Xueliang Zhao, Wei Wu, Jian Guan, and Lingpeng Kong. 2025. Promptcot: Synthesizing olympiad-level problems for mathematical reasoning in large language models. *arXiv preprint arXiv:2503.02324*.

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-o1: Towards open reasoning models for open-ended solutions. *Preprint*, arXiv:2411.14405.

Jiawei Zheng, Hanghai Hong, Feiyan Liu, 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*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and 1 others. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

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 Details of Multi-domain Test Sets

A.1 Composition and Scope of Multi-domain Test Sets

Our multi-domain test sets are designed to ensure broad coverage, clean splits, and reproducibility across language pairs. For En⇔De, we adopt well-partitioned and publicly released test sets from WMT, covering diverse domains such as news, IT, medical, legal, and subtitles. For Zh⇔En, we use publicly available and clearly defined corpora including CAMT, UM-Corpus, and CommonMT, which provide explicit training/test splits and standardized evaluation protocols. This design allows us to assess model performance on a wide spectrum of domain conditions—ranging from open-domain conversational text to terminology-intensive content—while minimizing potential data leakage and ensuring comparability across systems.

A.2 Data Sources

For German⇔English, the evaluation spans 11 domains. Five domains—Medical, Law, IT, Koran, and Subtitles—are sourced from the publicly available Multi-Domain dataset (Aharoni and Goldberg, 2020). The Mixed domain is drawn from the WMT22 General Machine Translation Task (Kocmi et al., 2022). Its subdomains include News, Social, E-commerce, and Conversation. Biomedical data is derived from the WMT18 and WMT19 Biomedical Machine Translation Task (Neves et al., 2018; Bawden et al., 2019).

For Chinese⇔English, the evaluation covers 12 domains. The CommonSense and Culture domains are taken from the CommonMT (He et al., 2020) and CAMT (Yao et al., 2024) datasets, respectively, while the Literary domain comes from the WMT23 Literary Machine Translation Task (Wang et al., 2023b). The general-purpose Mixed domain, along with its subdivisions (News, Social, E-commerce, and Conversation), is again sourced from the WMT22 General Machine Translation Task (Kocmi et al., 2022). Biomedical data is also collected from WMT18 (Neves et al., 2018) and WMT19 (Bawden et al., 2019) Biomedical Machine Translation Task. Additionally, the Laws, News, Science, and Subtitles domains are sourced from the UM-Corpus (Tian et al., 2014).

A.3 Data Statistics

We compile a collection of publicly available translation test sets covering diverse domains and four

Test set	Num	Test set	Num
Conversation	349	Social	491
E-commerce	518	News (WMT22)	499
Bio (WMT18)	212	Bio (WMT19)	243
Commonsense	1200	Literary	3038
Terminology	1638	Mixed (WMT22)	1875

Table 10: Test sets and the number of samples for Zh⇒En translation tasks. Bio denotes the Biomedical domain.

Test set	Num	Test set	Num
Culture	778	Laws	456
Subtitles	597	Science	503
Conversation	484	Social	511
E-commerce	530	News (WMT22)	511
Bio (WMT18)	244	Bio (WMT19)	224
News (UM)	1500	Mixed	2037

Table 11: Test sets and the number of samples for En⇒Zh translation tasks.

Test set	Num	Test set	Num
Bio (WMT18)	217	Bio (WMT19)	373
Social	515	E-commerce	501
News (WMT22)	506	Conversation	462
IT	2000	Law	2000
Koran	2000	Subtitles	2000
Medical	2000	Terminology	2955
Mixed	1984		

Table 12: Test sets and the number of samples for De⇒En translation tasks.

Test set	Num	Test set	Num
Bio (WMT18)	220	Bio (WMT19)	435
News (WMT22)	511	E-commerce	530
Social	512	Conversation	484
Mixed	2037		

Table 13: Test sets and the number of samples for En⇒De translation tasks.

translation directions: De⇒En, En⇒De, En⇒Zh, and Zh⇒En. The datasets span general-purpose and domain-specific content, including biomedical, legal, IT, conversational, and literary texts, etc. Sample sizes for each domain are reported in Tables 10 to 13, with no additional filtering or augmentation applied.

B Evaluation Metric Details

B.1 Automatic Metrics

We use three automatic evaluation metrics:

- **BLEU**[1] (Post, 2018), which evaluates surface-level n-gram overlap between system output and reference translations.
- **COMET**[2] (Rei et al., 2022a), a reference-based semantic metric trained on human quality judgments.
- **CometKiwi**[3] (Rei et al., 2022b), a reference-free version of COMET, useful when reference quality is poor or unavailable.

B.2 MQM Evaluation Protocol

To evaluate translation quality across domains and systems, we adopt an enhanced MQM hierarchy tailored to the characteristics of LLMs. Based on the official MQM taxonomy, we introduce additional error types frequently observed in LLM outputs—such as hallucination, semantic repetition, and cross-domain confusion. Our final schema spans seven dimensions: Accuracy, Fluency, Style, Terminology, Others, Source Error, and Non-translation Error, covering 25 fine-grained error categories (Table 14).

For consistent and scalable annotation, we employ DeepSeek-V3 as the automatic scoring model across all evaluations. It is applied uniformly to both traditional and LRMs to ensure fair comparison. The model follows a structured prompt designed to mimic human assessment while enforcing strict formatting and error attribution rules. The full prompt is shown in Figure 3.

C Evaluation Results

We report detailed automatic evaluation results across all domains and translation directions using BLEU, COMET, and CometKiwi. These results are summarized in Tables 15 through 25, covering all test sets introduced in our main experiments. Each table corresponds to a specific dataset or language direction, facilitating fine-grained comparison across models.

D API Details

The OpenAI, DeepSeek, and Gemini models used in this study are accessed via the following APIs: gpt-4o-2024-11-20, o1-2024-12-17, and o3-mini-2025-01-31 for OpenAI; deepseek-chat-2024-12-26 and deepseek-reasoner for DeepSeek; and gemini-2.0-flash and gemini-2.0-flash-thinking-

[1] https://github.com/mjpost/sacrebleu
[2] Unbabel/wmt22-comet-da
[3] Unbabel/wmt22-cometkiwi-da

Category	Error Type	Description
Accuracy	Mistranslation	Inaccurate translation causing semantic distortion.
	Addition	Adding extra information or emotions not in the source.
	Under-translation	Failure to fully convey cultural or contextual nuances.
	Omission	Unintentional exclusion of content from the source text.
	Untranslated	Retaining source text without translation.
	Hallucination	Generating content unrelated to the source text.
	Off-target Translation	Translation misalignment caused by ambiguous input.
	Contradiction	Contradicting itself or the source text.
Fluency	Grammar	Errors in sentence structure or syntax.
	Punctuation	Incorrect use of punctuation marks.
	Spelling	Misspelling of words.
	Semantic Repetition	Unnecessary repetition of words or phrases.
	Logical Incoherence	Lack of logical flow or coherence in translation.
Style	Awkward Expression	Stilted or unnatural phrasing in the target language.
	Unidiomatic Usage	Literal translation causes unnatural wording or grammar.
	Style Inconsistency	Inconsistent stylistic choices within the translation.
	Over-localization	Excessive cultural adaptation leading to distortion.
Terminology	Terminology Inconsistency	Inconsistent translation of the same term.
	Terminology Misuse	Use of incorrect or inappropriate terms for the domain.
	Cross-domain Confusion	Misuse of terms due to domain shifts.
	Incorrect Unit Conversion	Errors in unit conversion (e.g., metric to imperial).
	Formatting	Errors in domain-specific formats (e.g., legal, medical).
Others		Any other errors not covered in the above categories.
Source Error		Errors present in the Source text itself.
Non-translation Error		Translation is unassessable and unrelated to the Source.

Table 14: MQM Hierarchy.

Model	Culture		
	BLEU	COMET	KIWI
NLLB-200-3.3B	25.86	76.75	76.44
GoogleTrans	27.18	72.90	73.41
GPT-4o	38.65	85.24	83.14
DeepSeek-V3	40.27	85.46	83.21
Gemini-2.0-Flash	39.11	85.02	82.94
OpenAI-o1	34.57	85.06	84.25
DeepSeek-R1	35.88	85.45	83.14
OpenAI-o3-mini	34.05	84.36	84.25
Gemini-2.0-Flash-Thinking	36.94	84.63	83.68

Table 15: Results on the En⇒Zh Culture-MT dataset.

exp-2025-01-21 for Gemini.

E Prompt

The prompt for calculating MQM scores is provided in Figure 3. The prompt for evaluating translation difficulty is shown in the Figure 4.

F Case Study

To better understand the reasoning capacity of LRMs in MDMT, we conduct a qualitative case study using four representative examples drawn from different domains (see Figures 5, 6, and 7). These cases highlight both *what* needs to be reasoned and *how* reasoning manifests in the translation process.

F.1 What to reason.

Following our taxonomy of MDMT challenges, we identify three core reasoning demands: (1) **domain disambiguation**, (2) **terminology resolution**, and (3) **stylistic and structural alignment**. For instance, in the legal-domain example from Figure 5, the source phrase *"die Kreuz-Sieben"* requires contextual interpretation. While the literal surface form may suggest "cross-seven," DeepSeek-R1 correctly infers its domain as card games and outputs *"the seven of clubs"*, showing successful domain-triggered lexical reasoning.

In the cultural idiom example from Figure 6, the phrase *"下海经商"* is situated in the reform-era socio-economic context. DeepSeek-R1 renders this as *"ventured into business"*, demonstrating accurate idiomatic grounding and historical awareness. In Figure 7, the model handles the syntactically ambiguous sentence *"喜欢干净的小孩的人很多"* with correct attachment and referent mapping, outputting *"There are many people who like clean children"*. This requires internal logical inference

Model	Contextual Syntactic			Contextless Syntactic			Lexical			Average		
	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI
NLLB-200-3.3B	23.22	79.90	77.24	17.86	78.61	72.99	25.80	77.05	75.20	22.30	78.52	75.14
GoogleTrans	37.58	84.50	79.93	35.05	84.28	75.94	40.40	82.11	77.30	37.68	83.63	77.72
GPT-4o	31.42	85.11	81.88	32.82	85.24	78.35	33.89	85.41	79.83	32.71	85.25	80.02
DeepSeek-V3	30.19	84.90	81.91	33.53	85.58	78.76	35.09	85.52	79.69	32.94	85.33	80.12
Gemini-2.0-Flash	31.15	84.86	81.79	31.69	85.20	77.99	32.13	84.03	78.84	31.66	84.70	79.54
OpenAI-o1	29.80	84.74	80.66	28.04	85.14	79.48	30.40	84.00	79.70	28.50	84.96	81.11
DeepSeek-R1	26.28	84.54	82.07	30.00	84.96	78.62	30.40	84.00	79.70	28.89	84.50	80.13
OpenAI-o3-mini	28.55	84.57	79.19	27.18	84.43	82.97	29.43	83.82	80.87	28.39	84.27	81.01
Gemini-2.0-Flash-Thinking	31.54	84.61	81.96	31.60	84.34	77.66	33.49	84.59	79.95	32.21	84.51	79.86

Table 16: Results on the Zh⇒En Commonsense-MT dataset.

Model	News			Laws			Science			Subtitles		
	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI
NLLB-200-3.3B	18.30	75.02	68.66	44.72	89.98	85.35	19.91	78.26	73.93	16.90	75.36	71.78
GoogleTrans	26.39	77.55	76.69	29.94	83.81	81.69	25.45	79.36	77.61	21.92	75.39	75.63
GPT-4o	34.58	86.22	85.02	58.00	91.50	86.65	32.09	87.34	84.84	24.91	81.91	81.93
DeepSeek-V3	35.86	86.94	85.44	80.33	93.36	86.55	33.54	88.28	85.46	23.66	82.44	82.69
Gemini-2.0-Flash	36.39	86.37	85.14	66.71	92.23	86.49	33.47	87.69	85.24	24.64	81.91	82.33
OpenAI-o1	33.93	86.61	85.47	55.48	91.61	86.73	32.32	87.92	85.56	22.96	82.31	82.60
DeepSeek-R1	33.35	86.67	85.40	78.81	93.16	86.40	31.82	88.08	85.47	20.50	82.10	82.84
OpenAI-o3-mini	32.90	86.29	85.35	51.18	91.15	86.92	31.38	87.42	85.24	22.95	81.90	82.40
Gemini-2.0-Flash-Thinking	33.90	86.02	84.79	66.99	92.12	86.40	31.98	87.47	85.06	23.83	80.90	81.22

Table 17: Results on the En⇒Zh UM-Corpus dataset.

over noun phrase dependencies.

F.2 How to reason.

Across all examples, LRMs exhibit multi-step reasoning behavior. Rather than direct string-level translation, the model performs implicit sub-tasks: (1) parsing semantic and syntactic structure, (2) inferring latent domain cues, and (3) resolving lexical choices consistent with domain and register. In the legal example (Figure 5), DeepSeek-R1 successfully activates the domain schema for card games without explicit prompting, choosing standard terminology. In the idiom case (Figure 6), its lexical choice is informed by both temporal reference (*"the 1990s"*) and socio-political context, suggesting reasoning over historical semantics. Finally, the model's treatment of syntactic attachment in Figure 7 reflects internal consistency checks and structure-sensitive generation.

These cases illustrate how LRMs like DeepSeek-R1 leverage implicit reasoning over domain, lexis, and structure to produce more accurate and contextually grounded translations.

Model	Conversation			E-commerce			News			Social			Average		
	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI
NLLB-200-3.3B	21.03	76.81	74.35	13.47	72.62	70.48	19.81	78.57	77.38	23.12	77.65	75.34	19.36	76.41	74.39
GoogleTrans	25.73	81.10	76.95	21.32	78.60	75.47	29.07	82.13	79.67	36.57	82.25	78.28	28.17	81.02	77.59
GPT-4o	31.28	84.86	79.83	19.94	81.75	78.24	25.21	83.04	81.10	27.29	83.77	80.64	25.93	83.36	79.95
DeepSeek-V3	30.26	84.96	79.67	19.47	81.55	78.32	25.77	83.00	81.18	27.55	83.56	80.64	25.76	83.27	79.95
Gemini-2.0-Flash	32.79	84.46	79.19	19.97	81.54	77.90	26.62	83.04	80.93	29.25	83.24	80.16	27.16	83.07	79.55
OpenAI-o1	29.52	84.95	80.90	17.82	81.81	79.16	22.52	82.96	82.10	23.85	83.20	81.39	23.43	83.23	80.89
DeepSeek-R1	26.82	84.61	79.40	15.63	80.98	77.87	21.21	82.34	80.95	22.18	82.69	80.05	21.46	82.66	79.57
OpenAI-o3-mini	30.57	84.25	79.48	18.80	81.29	78.23	23.68	82.88	81.16	26.98	83.34	80.61	25.01	82.94	79.87
Gemini-2.0-Flash-Thinking	30.98	84.17	79.28	19.06	81.21	77.76	25.50	82.90	80.85	27.54	83.04	80.13	25.77	82.83	79.50

Table 18: Results on the Zh⇒En WMT22 General Machine Translation task.

Model	Conversation			E-commerce			News			Social			Average		
	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI
NLLB-200-3.3B	29.57	82.68	76.79	31.62	81.98	76.55	29.70	78.62	70.15	21.38	74.59	68.09	28.07	79.47	72.90
GoogleTrans	33.73	84.27	77.86	36.82	82.64	78.54	39.98	79.06	75.20	30.51	76.42	74.68	35.26	80.60	76.57
GPT-4o	45.63	89.44	82.67	43.82	88.82	83.52	51.20	88.31	83.41	38.60	84.57	81.33	44.81	87.78	82.73
DeepSeek-V3	43.49	89.89	82.94	42.84	89.00	83.79	50.28	88.75	83.76	37.59	85.27	82.24	43.55	88.23	83.18
Gemini-2.0-Flash	44.59	90.19	83.01	42.95	88.90	83.47	52.12	88.46	83.28	38.11	84.71	81.59	44.44	88.06	82.84
OpenAI-o1	42.24	89.86	84.13	39.22	89.06	84.59	45.88	88.47	84.67	33.41	85.07	82.98	40.19	88.11	84.09
DeepSeek-R1	38.00	88.74	82.74	34.48	88.35	83.48	42.56	88.08	83.58	29.43	84.21	82.01	36.12	87.34	82.95
OpenAI-o3-mini	43.24	89.02	82.44	41.54	88.51	83.41	47.06	87.53	82.94	36.38	83.99	81.27	42.05	87.26	82.52
Gemini-2.0-Flash-Thinking	44.31	89.71	82.66	42.39	88.73	83.31	49.54	88.23	83.26	36.35	83.86	81.08	43.15	87.63	82.58

Table 19: Results on the En⇒Zh WMT22 General Machine Translation task.

Model	Conversation			E-commerce			News			Social			Average		
	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI
NLLB-200-3.3B	35.09	86.37	81.94	33.45	87.16	83.57	32.78	85.24	82.76	32.48	82.94	81.57	33.45	85.43	82.46
GoogleTrans	32.63	83.73	79.26	29.89	82.83	79.37	28.04	79.37	77.51	28.57	77.58	77.51	29.78	80.88	78.41
GPT-4o	40.65	89.88	83.72	38.03	88.80	84.02	36.70	87.35	84.00	34.06	85.12	82.63	37.36	87.79	83.59
DeepSeek-V3	39.58	89.33	83.50	37.42	88.69	84.14	36.48	87.94	84.76	35.08	85.22	82.91	37.14	87.79	83.83
Gemini-2.0-Flash	37.88	88.46	83.36	37.47	88.98	84.09	35.88	87.69	84.63	34.02	85.39	82.84	36.31	87.63	83.73
OpenAI-o1	38.71	90.19	85.55	36.25	88.73	85.79	33.90	88.17	86.69	32.18	85.73	84.54	35.26	88.21	85.64
DeepSeek-R1	36.17	89.11	83.60	34.70	88.73	84.27	33.96	87.92	84.87	32.39	85.22	82.85	34.31	87.75	83.90
OpenAI-o3-mini	35.35	88.05	83.21	36.71	88.28	84.01	34.01	87.16	84.31	32.14	84.29	82.25	34.55	86.95	83.44
Gemini-2.0-Flash-Thinking	36.64	88.04	83.00	37.87	88.97	83.98	36.18	87.81	84.71	33.84	85.00	82.59	36.13	87.45	83.57

Table 20: Results on the En⇒De WMT22 General Machine Translation task.

Model	Conversation			E-commerce			News			Social			Average		
	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI
NLLB-200-3.3B	29.25	83.31	80.39	29.82	83.93	80.68	26.98	83.17	80.59	26.76	81.68	79.68	28.20	83.02	80.33
GoogleTrans	27.69	82.51	78.66	30.79	81.57	77.56	28.33	81.04	78.09	28.09	79.80	76.65	28.72	81.23	77.74
GPT-4o	37.14	86.07	81.79	34.87	86.15	81.37	31.74	85.49	82.14	32.35	85.07	81.27	34.02	85.69	81.64
DeepSeek-V3	36.52	86.24	81.84	34.56	86.04	81.46	30.76	85.52	82.47	31.65	84.93	81.43	33.37	85.68	81.80
Gemini-2.0-Flash	36.76	86.05	81.92	34.30	86.02	81.23	31.28	85.64	82.35	31.57	84.88	81.37	33.48	85.65	81.72
OpenAI-o1	34.73	86.49	83.40	31.91	85.95	82.65	30.35	85.78	83.93	30.36	85.22	82.93	31.84	85.86	83.23
DeepSeek-R1	35.20	86.06	81.96	32.11	85.75	81.39	29.59	85.78	82.66	30.04	84.92	81.45	31.73	85.63	81.86
OpenAI-o3-mini	36.16	85.66	81.54	34.20	85.59	81.26	31.07	84.71	81.30	32.02	84.66	81.00	33.36	85.16	81.28
Gemini-2.0-Flash-Thinking	36.35	86.05	81.79	33.05	85.77	81.14	31.59	85.83	82.42	31.48	84.76	81.28	33.12	85.60	81.66

Table 21: Results on the De⇒En WMT22 General Machine Translation task.

Model	WMTBio18 En⇒Zh			WMTBio18 Zh⇒En			WMTBio18 En⇒De			WMTBio18 De⇒En			Average		
	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI
NLLB-200-3.3B	26.52	81.50	80.09	20.79	81.62	82.22	21.30	85.11	83.70	27.88	84.35	82.13	24.62	83.15	82.03
GoogleTrans	30.11	81.74	79.74	24.27	82.85	82.59	17.90	82.26	81.10	25.25	81.23	79.43	24.88	82.02	80.72
GPT-4o	40.86	86.66	84.61	25.04	84.33	84.02	24.44	86.68	84.97	29.31	85.06	82.61	29.91	85.68	84.05
DeepSeek-V3	40.58	86.67	84.66	25.73	84.42	83.95	24.75	86.85	84.96	29.37	85.01	82.78	30.11	85.74	84.09
Gemini-2.0-Flash	40.83	86.87	84.62	26.32	84.50	83.74	26.47	86.72	84.66	30.69	85.10	82.22	31.08	85.80	83.81
OpenAI-o1	38.84	86.63	85.52	24.85	84.30	84.88	32.29	86.96	86.68	28.42	85.08	84.05	31.35	85.74	85.38
DeepSeek-R1	38.43	86.68	84.45	24.34	84.22	83.70	25.14	86.94	85.01	29.34	85.04	82.79	29.31	85.72	83.99
OpenAI-o3-mini	39.39	86.15	84.36	25.09	83.99	83.77	22.96	86.39	84.73	29.06	84.70	82.77	29.12	85.31	83.91
Gemini-2.0-Flash-Thinking	39.90	86.36	84.35	25.54	84.29	83.90	25.78	86.75	84.72	30.57	84.98	82.86	30.45	85.59	83.95
Model	WMTBio19 En⇒Zh			WMTBio19 Zh⇒En			WMTBio19 En⇒De			WMTBio19 De⇒En			Average		
	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI
ALMA-13B-R	26.02	85.00	82.83	25.90	83.50	82.44	23.22	83.26	82.74	35.15	86.73	82.00	27.07	84.62	82.50
TowerInstruct-13B-v0.1	32.41	85.30	82.81	31.32	83.76	82.24	30.33	85.49	83.39	36.34	86.74	81.92	32.10	85.32	82.59
GemmaX2-28-9B-v0.1	37.06	86.32	84.21	33.95	84.70	82.60	33.00	86.43	83.51	36.95	87.02	81.93	35.24	86.12	83.56
NLLB-200-3.3B	25.05	80.93	80.05	26.91	82.33	81.26	29.47	81.18	82.44	31.76	76.91	75.30	28.30	80.34	79.76
GoogleTrans	28.37	81.00	80.16	31.28	83.41	81.88	26.15	79.43	80.88	30.83	81.71	78.23	29.16	81.39	80.29
GPT-4o	37.58	86.75	84.53	31.41	84.42	82.72	30.49	86.29	83.55	37.25	86.91	82.35	34.18	86.09	83.29
DeepSeek-V3	38.54	86.93	84.59	30.75	84.68	83.10	34.14	86.47	83.91	37.63	87.66	82.29	35.26	86.44	83.48
Gemini-2.0-Flash	38.89	86.91	84.48	32.59	84.50	82.86	34.59	87.96	83.96	37.87	86.71	82.24	35.98	86.52	83.39
OpenAI-o1	37.01	86.97	85.42	28.15	84.53	83.50	32.29	83.52	85.88	36.62	84.93	83.82	33.52	85.00	84.66
DeepSeek-R1	37.23	87.05	84.27	28.58	84.31	82.60	32.56	83.80	83.98	36.19	84.84	82.28	33.64	85.00	83.28
OpenAI-o3-mini	35.62	86.07	84.22	30.59	84.04	82.91	31.55	86.16	83.86	36.92	86.48	82.19	33.67	85.69	83.29
Gemini-2.0-Flash-Thinking	38.48	86.43	84.20	31.78	84.48	82.88	33.55	86.82	83.89	36.74	84.57	82.25	35.14	85.57	83.31

Table 22: Results on the Zh⇔En and De⇔En Biomedical Machine Translation Task from WMT18 and WMT19 (WMTBio18 and WMTBio19).

Model	Medical			Law			IT			Koran			Subtitles			Average		
	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI
NLLB-200-3.3B	40.38	76.99	74.41	24.43	75.23	76.38	35.02	78.21	76.54	29.52	75.91	72.44	32.26	79.97	80.34	38.11	78.21	76.54
GoogleTrans	33.35	80.16	78.79	10.03	64.16	71.72	24.27	75.93	76.51	27.28	79.28	75.71	24.58	76.91	79.16	24.08	75.93	76.51
GPT-4o	41.89	84.23	81.89	39.00	85.06	82.38	37.23	83.52	78.37	17.65	75.04	80.64	29.61	80.62	82.25	33.08	81.69	81.11
DeepSeek-V3	41.43	84.15	81.90	38.82	84.94	82.29	38.16	83.90	78.86	17.94	74.91	80.33	29.70	80.68	82.26	33.21	81.72	81.13
Gemini-2.0-Flash	44.39	84.49	81.76	46.15	85.57	82.30	37.93	83.20	78.49	19.70	74.97	79.71	30.39	80.74	82.19	35.71	81.79	80.89
OpenAI-o1	40.44	83.88	83.29	40.74	85.03	83.53	36.40	83.20	79.75	16.76	75.03	82.02	28.91	80.79	83.65	32.65	81.59	82.45
DeepSeek-R1	40.23	83.82	81.81	39.16	84.97	82.35	36.66	83.49	78.68	17.05	74.76	80.63	28.82	80.65	82.22	32.38	81.54	81.14
OpenAI-o3-mini	39.82	83.60	81.79	34.29	84.18	81.96	37.13	82.63	78.20	15.52	73.54	80.19	28.39	80.06	82.06	31.03	80.80	80.84
Gemini-2.0-Flash-Thinking	43.36	84.14	81.70	44.51	85.10	82.13	37.26	82.94	78.40	19.57	75.03	79.79	30.38	80.78	82.20	35.02	81.60	80.84

Table 23: Results on the Dn⇒En Multi-Domain dataset.

Model	Test-1			Test-2			Valid-1			Valid-2			Average		
	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI
NLLB-200-3.3B	11.14	66.97	66.78	13.66	66.63	64.88	9.90	67.48	68.56	14.12	68.57	67.47	11.14	66.97	66.78
GoogleTrans	20.43	76.20	74.63	27.54	76.65	73.65	15.25	76.26	75.63	26.72	78.20	74.94	20.98	76.11	74.70
GPT-4o	20.49	78.44	76.50	16.12	77.55	77.82	21.08	78.91	77.31	12.49	75.38	76.91	17.54	77.57	77.13
DeepSeek-V3	15.66	77.47	77.84	16.25	77.14	77.49	19.57	78.94	77.48	11.87	75.28	77.05	16.81	77.48	77.17
Gemini-2.0-Flash	22.11	78.25	75.94	16.25	77.14	77.49	21.42	78.70	76.70	12.70	75.11	76.34	18.12	77.30	76.62
OpenAI-o1	19.05	78.85	78.00	16.04	77.76	78.66	18.03	78.13	76.93	12.65	75.52	77.44	16.44	77.56	77.76
DeepSeek-R1	15.34	77.29	75.72	12.68	77.04	77.23	14.35	77.67	76.59	10.09	74.55	76.33	13.12	76.64	76.47
OpenAI-o3-mini	20.63	77.77	76.02	15.57	76.69	77.38	21.11	78.35	76.97	11.74	74.69	76.17	17.26	76.88	76.63
Gemini-2.0-Flash-Thinking	21.69	78.78	76.71	15.75	76.98	77.43	21.14	78.17	75.90	12.48	74.75	76.27	17.77	77.17	76.58

Table 24: Results on the Zh⇒En GuoFeng Webnovel dataset (Literary domain).

Model	WMT22 En⇒Zh			WMT22 Zh⇒En			WMT22 En⇒De			WMT22 De⇒En			Average		
	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI	BLEU	COMET	KIWI
NLLB-200-3.3B	28.56	79.45	72.88	19.34	76.35	74.36	33.45	85.43	82.47	28.14	83.01	80.33	28.07	81.06	77.51
GoogleTrans	36.50	80.57	76.57	29.41	81.00	77.63	29.48	80.86	78.41	28.92	81.19	77.71	31.58	80.91	77.58
DeepSeek-V3	44.52	88.21	83.19	25.21	83.11	79.95	36.89	87.78	83.83	33.06	85.66	81.80	34.92	86.19	82.19
GPT-4o	45.64	87.77	82.74	25.11	83.22	79.95	37.07	87.77	83.60	33.71	85.68	81.64	35.38	86.11	81.98
Gemini-2.0-Flash	45.50	88.04	82.84	26.12	82.94	79.56	36.16	87.63	83.74	33.20	85.63	81.71	35.25	86.06	81.96
OpenAI-o1	40.84	88.10	84.10	22.42	83.08	80.87	34.87	88.18	85.64	31.53	85.84	83.22	32.42	86.30	83.46
DeepSeek-R1	36.88	87.33	82.96	20.68	82.48	79.57	34.14	87.73	83.91	31.37	85.61	81.86	30.77	85.79	82.07
OpenAI-o3-mini	42.67	87.25	82.52	24.09	82.82	79.89	34.46	86.94	83.45	33.08	85.14	81.27	33.58	85.54	81.78
Gemini-2.0-Flash-Thinking	44.04	87.61	82.58	24.83	82.71	79.51	36.12	87.46	83.58	32.75	85.59	81.66	34.44	85.84	81.83

Table 25: Results on the Zh⇔En and De⇔En WMT22 General Machine Translation task.

You are a professional translation quality evaluator following the MQM (Multidimensional Quality Metrics) framework.
Task Instructions:

1. Compare the Prediction against both the Source and the Reference.
2. Identify up to five of the most serious Errors for each translation sentence, using the MQM error types listed below.
3. Assign exactly one severity level to each error.
4. Special handling for two specific error types:
   - **Source Error**: Errors present in the Source text itself.
   - **Non-translation Error**: Translation is unassessable and unrelated to the Source.
5. If no errors are found, return an empty JSON list [].
6. Only output a valid JSON object. Do not include any additional text, comments, or explanations.

MQM Error Types (See Table 14 for the complete hierarchy):

- Mistranslation: Incorrect translation that alters or distorts meaning.
- Addition: Insertion of information or emotion not present in the source.
- Under-translation: Partial omission of relevant or necessary information.
- Omission: Complete exclusion of content present in the source.
- Untranslated: Source text is copied without being translated.

Severity Levels:

- Minor: Slight impact on readability or style; meaning remains clear.
- Major: Significant impact on usability, comprehension, or meaning.

Output Format:

```
{
  "errors": [
    {
      "error_type": "Mistranslation",
      "severity": "Major",
      "explanation": "The correct translation should be ... "
    }
  ]
}
```

Source: `{source_text}`
Reference: `{reference_text}`
Prediction: `{prediction_text}`

Figure 3: Full prompt used to calculate MQM scores with DeepSeek-V3.

Your task is to assess the difficulty of translating a given {src_lang} sentence into {tgt_lang}. Please evaluate the difficulty based on the following criteria and output the result in JSON format, with the key "level":

1. Sentence complexity: Determine if the sentence is a simple sentence, a compound sentence, or includes subordinate clauses and other complex structures.
2. Vocabulary difficulty: Assess whether the sentence contains commonly used words or specialized terms or slang.
3. Grammar differences: Analyze if the sentence's grammatical structure is similar to or differs significantly from {tgt_lang}.
4. Contextual understanding: Consider whether understanding specific cultural contexts or background knowledge is necessary for accurate translation.

Provide a difficulty level (1-5), with 1 being the easiest and 5 being the most difficult. And output the difficulty level in the following JSON format:

```
{
    "level": "difficulty level"
}
```

Here is the {src_lang} sentence: {src_text}

Figure 4: Full prompt used for evaluating translation difficulty with DeepSeek-V3.

Source: Entfernen Sie die Kreuz-Sieben.	Source: Der Karren ist natürlich weg.
Reference: Remove the seven of clubs.	Reference: The cart is gone, of course.
DeepSeek-R1: Remove the seven of clubs.	DeepSeek-R1: The cart is gone, of course.
DeepSeek-V3: Remove the cross-seven.	DeepSeek-V3: The cart is naturally gone.
Reasoning: ...First, I need to understand the original sentence. ..."die Kreuz-Sieben" – "Kreuz" translates to "club" in the context of card suits (like in playing cards: clubs, spades, hearts, diamonds). "Sieben" is "seven." So "Kreuz-Sieben" would be the seven of clubs. Putting it together, the sentence is instructing someone to remove the seven of clubs. But I should check if there's any idiomatic expression or specific context where "Kreuz-Sieben" might mean something else. However, in most cases, especially in card games, it's straightforward. Now, translating the imperative: "Entfernen Sie" becomes "Remove" in English. Since it's formal, maybe "Please remove" to keep the politeness. But the original doesn't have a "bitte," so just "Remove the seven of clubs." Let me confirm once more. "Entfernen Sie die Kreuz-Sieben." → "Remove the seven of clubs." Yep, that's accurate.	Reasoning: Alright, let's tackle this translation. The German sentence is "Der Karren ist natürlich weg." First, I need to understand each word. "Der Karren" translates to "the cart" or "the wagon." Then "ist natürlich weg." "Ist" is "is," "natürlich" is "naturally" or "of course," and "weg" means "gone" or "away." So putting it together: "The cart is naturally gone." But that sounds a bit off in English. The phrase "natürlich" here is likely used in the sense of "of course" rather than "naturally." So maybe "The cart is of course gone." That makes more sense. Alternatively, "The cart is gone, of course." Which flows better in English. ...Maybe the latter is more conversational. "The cart is gone, of course." Yes, that sounds natural. So I'll go with that."

Figure 5: Case study of DeepSeek-R1 showing its reasoning process and final translation.

Source: 九十年代很多人下海经商。

Reference: In the 1990s, many people ventured into business.

DeepSeek-R1: In the 1990s, many people ventured into business.

DeepSeek-V3: In the 1990s, many people went into business.

Reasoning: 首先，我得准确理解原句的意思。"九十年代"应该是指1990年代，也就是the 1990s。这里的"下海"是一个习语，通常指离开原来的稳定工作去从商，尤其是在中国改革开放初期，很多人放弃铁饭碗投身私营经济。所以"下海经商"可以翻译为"ventured into business"或者"started their own businesses"。"很多人"就是many people。所以整个句子的大意是"In the 1990s, many people ventured into business." 或者更口语化的表达，比如"In the 90s, a lot of people started businesses."
First, I need to accurately understand the meaning of the original sentence. "九十年代" refers to the 1990s. The term "下海" is an idiom that usually means leaving a stable job to go into business, especially during the early years of China's reform and opening-up, when many gave up their iron rice bowls to join the private economy. So "下海经商" can be translated as "ventured into business" or "started their own businesses." "很多人" means "many people." Therefore, the overall sentence could be: "In the 1990s, many people ventured into business." Or, in a more colloquial form: "In the 90s, a lot of people started businesses."
接下来，我需要确保语法正确，时态合适。因为说的是过去的事情，用一般过去式。比如"ventured"或者"started"。
Next, I need to ensure that the grammar is correct and the tense is appropriate. Since this describes a past event, the simple past tense should be used, such as "ventured" or "started."
同时要注意"下海"这个动作，可能需要用更准确的动词来传达那种冒险的感觉。比如"ventured"比较贴切，因为它有冒险尝试的意思。
I also need to pay attention to the action of "下海," which may require a more precise verb to convey the sense of taking a risk. "Ventured" fits well here because it carries the meaning of trying something adventurous.
然后，确定正确的年代表达方式，是"the 1990s"还是"the 90s"。两者都可以，但前者更正式，后者更口语。根据原句的语气，可能前者更合适，因为原句没有明显的口语化用词...
Then, I should determine the correct way to express the decade: "the 1990s" or "the 90s." Both are acceptable, but the former is more formal, while the latter is more conversational. Based on the tone of the original sentence, the formal version may be more appropriate since there are no clearly colloquial expressions in the original...
检查一下有没有拼写错误，语法是否正确...
Lastly, check for any spelling mistakes and ensure grammatical correctness...

Figure 6: Case study of DeepSeek-R1 showing its reasoning process and final translation.

Source: 喜欢干净的小孩的人很多。

Reference: There are many people who like clean children.

DeepSeek-R1: There are many people who like clean children.

DeepSeek-V3: Many people like clean children.

Reasoning: 首先，我得仔细理解原句的意思。原句的主干是"人很多"，而"喜欢干净的小孩的"是修饰"人"的定语从句。所以结构上，这是一个存在很多人的情况，这些人喜欢干净的小孩。
First, I need to carefully understand the meaning of the original sentence. The main part of the sentence is "人很多" and "喜欢干净的小孩的" is a relative clause modifying "人" Structurally, this indicates a situation where there are many people, and these people like clean children.
接下来要考虑如何准确翻译这个定语结构。中文里的定语通常放在名词前面，但英语中如果定语较长，可能会使用关系从句或者介词结构。这里"喜欢干净的小孩的"可以翻译成"who like clean children"或者"who like children who are clean"。不过可能更自然的方式是简化，比如"people who like neat children"或者"people who like clean kids"。
Next, I need to consider how to translate this attributive structure accurately. In Chinese, attributive clauses usually precede the noun, but in English, if the attributive clause is long, it may use a relative clause or a prepositional structure. Here, "喜欢干净的小孩的" can be translated as "who like clean children" or "who like children who are clean." However, a more natural way might be to simplify it, such as "people who like neat children" or "people who like clean kids."
再考虑句子的结构是否需要调整。原句的主语是"人"，而"很多"作为谓语，可能英语中更适合用存在句型。例如："There are many people who..."这样的结构更自然，而不是直译成"The people who... are many."
Next, I should consider whether the sentence structure needs adjustment. In the original sentence, the subject is "人" and "很多" acts as the predicate. In English, it may be more appropriate to use an existential construction. For example, "There are many people who..." is more natural than directly translating it as "The people who... are many."
最后，检查语法是否正确，确保关系代词的使用正确，主谓一致，以及句子的流畅性。
Finally, I need to check whether the grammar is correct, ensuring the proper use of relative pronouns, subject-verb agreement, and the fluency of the sentence.

Figure 7: Case study of DeepSeek-R1 showing its reasoning process and final translation.