Reasoning for Translation: Comparative Analysis of Chain-of-Thought and Tree-of-Thought Prompting for LLM Translation

Anonymous ACL submission

Abstract

As Large Language Models (LLMs) continue to advance in capability, prompt engineering has emerged as a crucial method for optimizing their performance on specialized tasks. While prompting strategies like Zero-shot, Few-shot, Chain-of-Thought, and Tree-of-Thought have demonstrated significant improvements in reasoning tasks, their application to machine translation has received comparatively less attention. This paper systematically evaluates these prompting techniques across diverse language pairs and domains, measuring their effect on translation quality. Our findings reveal substantial performance variations between prompting methods, with certain strategies offering consistent improvements for specific language directions and complexity levels. These results provide valuable insights for developing more effective LLM-based translation systems without requiring model fine-tuning and complement existing works in the field.

1 Introduction

Large Language Models (LLMs) (Brown et al., 2020; OpenAI et al., 2024) have revolutionized Natural Language Processing, offering new capabilities for machine translation (MT) that challenge traditional paradigms. While conventional neural machine translation (NMT) systems (Bahdanau et al., 2016; Vaswani et al., 2017) depend on extensive supervised training with bilingual datasets, LLMs demonstrate impressive translation abilities that can be enhanced through strategic prompting rather than task-specific fine-tuning (Zhang et al., 2023). These prompting techniques—which have already transformed performance in reasoning (Wei et al., 2023), question-answering (Kojima et al., 2023), and mathematical problem-solving tasks (Yao et al., 2023)—represent a promising but understudied approach for translation. As organizations increasingly deploy LLMs for cross-lingual communication (Jiao et al., 2023), understanding how different prompting strategies affect translation quality across language pairs becomes essential for both practical applications and theoretical advancement of the field.

2 Related Works

2.1 LLMs for Machine Translation

Large Language Models (LLMs) (Minaee et al., 2024; Raiaan et al., 2024; Zhao et al., 2025; Brown et al., 2020) such as GPT-4 (OpenAI et al., 2024), Llama 3.3 (Grattafiori et al., 2024), Claude (Enis and Hopkins, 2024), and Qwen (Qwen et al., 2025) have demonstrated significant translation capabilities without translation-specific architectures. These models leverage their pre-training on vast multilingual corpora to perform cross-lingual tasks effectively (Lin et al., 2022; Ahuja et al., 2023; Zhu et al., 2024). Studies by (Jiao et al., 2023), (Coleman et al., 2024), and (Zhang et al., 2023) show LLMs can match specialized translation systems for certain language pairs, with particular advantages in domain adaptation and context handling (Zhang et al., 2025; Chen et al., 2022; Briva-Iglesias et al., 2024). LLMs excel at incorporating contextual information and maintaining semantic consistency across languages (Zhu et al., 2024; Garcia et al., 2023), though their performance varies substantially across language pairs (Sanh et al., 2022; Zhang et al., 2023). High-resource languages typically benefit from better representation in pre-training data (Kudugunta et al., 2023; Team et al., 2022), while low-resource languages often present ongoing challenges (Ahuja et al., 2023; Huang et al., 2023; Ghazvininejad et al., 2023). Unlike specialized translation models that require extensive fine-tuning for optimal results, LLMs can be adapted for translation tasks through prompt engineering techniques (Wei et al., 2023; Zhou et al., 2023; Liu et al., 2022), offering flexibility without the computational cost of retraining. However, challenges remain in optimizing these prompting approaches (Yao et al., 2023; Zhang et al., 2024), ensuring consistent quality across diverse language combinations (Zhu et al., 2024; Xie et al., 2023), and addressing the computational demands of inference with large models (Xia et al., 2024; Bapna and Firat, 2019).

2.2 Prompting Strategies for Translation

Prompting strategies fundamentally shape how LLMs approach translation tasks, offering different trade-offs between simplicity, performance, and computational efficiency. We examine four major prompting paradigms and their applications to machine translation.

2.2.1 Zero-shot & Few-shot prompting

Zero-shot prompting leverages an LLM's pretrained knowledge to perform translations without any task-specific examples (Brown et al., 2020). This approach relies entirely on the model's existing parameters, making its effectiveness heavily dependent on the language pair's representation in the pre-training corpus (Vilar et al., 2023). While effective for high-resource languages, zero-shot translation often falters with idiomatic expressions, rare vocabulary, and specialized terminology (Jiao et al., 2023).

Few-shot prompting aims to enhance translation quality by incorporating example translations directly in the prompt (Brown et al., 2020), as illustrated in Table 1. These in-context examples allow the model to recognize translation patterns specific to the current task, improving both accuracy and fluency (Tan et al., 2022). The effectiveness of few-shot prompting depends critically on three factors: (1) the quality of provided examples, (2) their diversity across linguistic constructions, and (3) their relevance to the target domain.

2.2.2 Chain-of-Thought & Tree-of-Thought prompting

While zero-shot and few-shot approaches provide direct translation, more sophisticated reasoning-based prompting techniques have emerged to address complex translation challenges. Chain-of-Thought (CoT) prompting (Wei et al., 2023) breaks down complex reasoning into intermediate steps, enabling LLMs to explicitly track grammatical transformations, handle idiomatic expressions, and maintain semantic consistency across languages. By decomposing the translation process, CoT can potentially improve handling of linguistic phenomena like long-range dependencies and structural divergences between languages.

Tree-of-Thought (ToT) prompting (Yao et al., 2023) extends this concept by enabling exploration of multiple translation candidates simultaneously. This approach allows the model to consider alternative phrasings, grammatical structures, or word choices before selecting the optimal translation path. Recent work by (Zhang et al., 2023) has begun exploring these advanced prompting strategies for translation, but comprehensive evaluation across diverse language pairs and LLM architectures remains limited.

2.3 Domain Adaptation & Noisy Texts MT

Domain adaptation in machine translation has been extensively studied, with comprehensive surveys provided by Chu and Wang (2018) and Saunders (2022). Previous work has explored various approaches, including nearest-neighbor methods (Martins et al., 2022), unsupervised learning techniques (Yang et al., 2018), and knowledge distillation (Wang et al., 2024). With the emergence of Large Language Models (LLMs) in machine translation, recent research has shifted toward multi-domain adaptation. Li et al. (2023) proposed a multi-task in-context learning approach, while Lu et al. (2024) introduced Chain-of-Dictionary prompting for low-resource language adaptation.

Handling noisy data remains a significant challenge in NLP. (Al Sharou et al., 2021) define noisy text characteristics, while (Yuan et al., 2024) leverage noisy labels to enhance LLM robustness. (Zheng and Saparov, 2023) improve multi-hop reasoning through noisy exemplars, and in machine translation, (Herold et al., 2022) explore noise detection for NMT. Prior work by (Bolding et al., 2023) employs LLMs for noise cleaning, and (Vogel, 2003) investigate the use of noisy bilingual datasets for NMT.

3 Methodology

3.1 Zero-Shot & Few-Shot Prompting for MT

For our experimental evaluation, we implemented zero-shot and few-shot prompting strategies as detailed in Table 1. For few-shot prompting, we carefully selected three representative examples per language pair, ensuring diversity in sentence length, grammatical structures, and vocabulary. Ex-

Zero-Shot Prompting [5]
`Translate the following sentence from [SRC] to [TGT]:` **`main text`**
Few-Shot Prompting (3-shot) [6]
`Translate the following sentence from [SRC] to [TGT]:` **`sample text 1`**   `Translate the following sentence from [SRC] to [TGT]:` **`sample text 2`**   `Translate the following sentence from [SRC] to [TGT]:` **`sample text 3`**   `Now, translate the following sentence from [SRC] to [TGT]:` **`main text`**

Table 1: Prompting templates for Zero-Shot and Few-Shot strategies in LLM-based machine translation.

ample selection was based on two criteria: (1) high-quality professional translations from parallel corpora, and (2) coverage of common linguistic phenomena in the target languages.

All prompts were kept consistent across experiments, with only the language pair identifiers (**[SRC] / [TGT]**) and text samples varying. This standardization ensures that performance differences can be attributed to the prompting strategy rather than prompt wording variations.

3.2 Advanced Prompting Techniques for MT

Beyond basic zero-shot and few-shot approaches, we investigate structured reasoning prompts that guide models through explicit translation processes. We evaluate two advanced techniques—Chain-of-Thought and Tree-of-Thought—across multiple translation tasks to assess their impact on accuracy, fluency, and contextual understanding.

3.2.1 CoT Prompting for MT

Chain-of-Thought (CoT) prompting (Wei et al., 2023) encourages step-by-step reasoning by decomposing complex tasks into intermediate steps. For translation, we formalize this as a process that transforms source text $x \in X$ into target text $y \in Y$ through a structured workflow of sequential operations.

Our implementation begins with a segmentation function $S : X \to \{x_1, x_2, ..., x_m\}$ that partitions complex input into manageable units. Each segment then undergoes processing through a translation engine T that implements a four-step reasoning

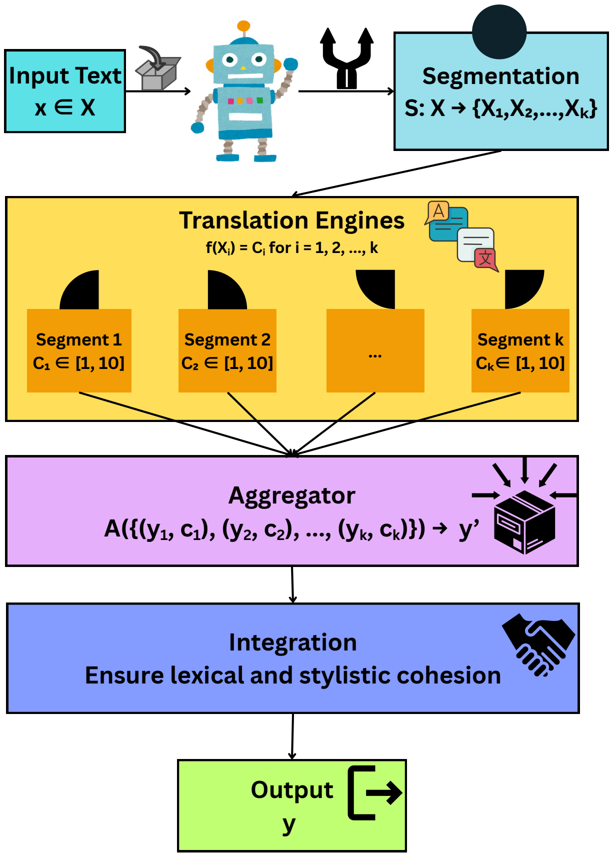

Figure 1: Chain-of-Thought (CoT) translation workflow featuring: (1) text segmentation, (2) sequential reasoning process (analysis, disambiguation, generation, verification), (3) confidence scoring, and (4) aggregation for cohesive output. This approach excels with complex syntactic structures and cultural nuances.

chain:

$$T(x_i) = f_{\text{verify}} \circ f_{\text{gen}} \circ f_{\text{disambig}} \circ f_{\text{analysis}}(x_i) \tag{1}$$

where f_{analysis} performs syntactic and semantic assessment, f_{disambig} resolves lexical ambiguities, f_{gen} produces the initial translation, and f_{verify} validates semantic equivalence. Each translated segment receives a confidence score $c_i \in [1, 10]$ based on the model's certainty.

The segments then flow through an aggregation function A that reconciles potential inconsistencies across segment boundaries:

$$A(\{(y_1, c_1), (y_2, c_2), ..., (y_m, c_m)\}) \to y' \tag{2}$$

A final integration step ensures lexical cohesion and stylistic consistency to produce the complete translation y.

Our experiments revealed mixed results across language pairs. CoT demonstrated statistically significant improvements ($p < 0.05$) for languages with substantial structural divergence from English (particularly Japanese and Chinese), but with modest overall gains. While the explicit reasoning steps sometimes effectively bridged linguistic gaps, they occasionally introduced error propagation or unnecessary verbosity that complicated the translation process.

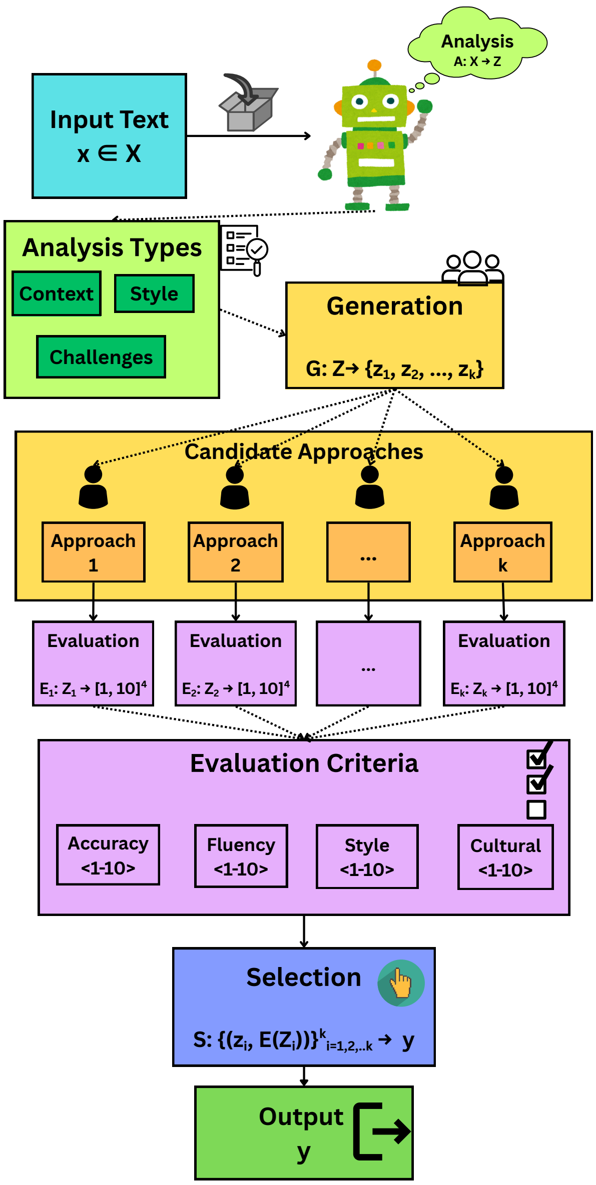

Figure 2: Tree-of-Thought (ToT) translation framework employing: (1) comprehensive text analysis, (2) parallel generation of multiple translation candidates, (3) multi-dimensional evaluation (accuracy, fluency, style, cultural appropriateness), and (4) weighted selection of optimal output. This approach excels with polysemous terms, idiomatic expressions, and culturally-specific content.

3.2.2 ToT Prompting for MT

Tree-of-Thought (ToT) prompting (Yao et al., 2023) extends the linear CoT approach by implementing a branching structure that explores multiple translation candidates simultaneously. Formally, ToT can be represented as a directed tree $T = (V, E)$ where nodes $v \in V$ correspond to translation states and edges $e \in E$ represent transitions between these states.

The process begins with a comprehensive text analysis function $A : X \to \mathcal{Z}$ that maps the source text $x \in X$ to a feature space $\mathcal{Z}$ capturing contextual dependencies, linguistic challenges, and stylistic elements. Unlike the sequential CoT approach, ToT then employs a branching generation function $G : \mathcal{Z} \to \{z_1, z_2, ..., z_k\}$ that produces k distinct translation candidates, where each z_i represents a different interpretation or rendering approach.

These candidates undergo multi-dimensional evaluation through a function $E : Z \to \mathbb{R}^4$ that maps each translation to a 4-tuple of scores $\langle s_{acc}, s_{flu}, s_{sty}, s_{cul} \rangle \in [1, 10]^4$ representing accuracy, fluency, stylistic fidelity, and cultural appropriateness, respectively. The final selection function $S : \{(z_i, E(z_i))\}_{i=1}^{k} \to y$ identifies the optimal translation by computing a weighted aggregate of these evaluation dimensions.

Our experiments demonstrate that ToT prompting outperforms baseline methods when handling polysemous terms, idiomatic expressions, and culturally-specific concepts. The approach shows particular strength in creative text domains where stylistic considerations are paramount, yielding improvements in human evaluation scores for literary translation tasks (will be described more careful in Section 4). However, this performance gain comes with increased computational costs of $O(k \cdot |x|)$ and prompt complexity that must be considered for practical applications.

3.3 Self-Guided Reasoning Promptings for MT

While previous sections examined structured reasoning across predefined prompting patterns, this section explores how LLMs can autonomously adapt to domain-specific content without explicit domain instructions (Wei et al., 2023; Yao et al., 2023). We formalize this approach as a two-phase translation process:

Standard Prompt
System: You are a machine translation system. **User:** Translate the following text from **[SRC]** to **[TGT]**: <input_text>
Domain-Specific Prompt (DSP)
System: You are a machine translation system that translates sentences in the **[DOMAIN]** domain. **User:** Translate the following text from **[SRC]** to **[TGT]**: <input_text>
Self-Guided CoT/ToT Prompt
System: You are a machine translation system. **User:** Translate from **[SRC]** to **[TGT]**: <input_text> **Domain Analysis:** • Extract specialized terminology and domain-specific jargon • Autonomously identify the domain (medical, legal, technical, etc.) • Determine appropriate register and stylistic conventions. Follow the template for translation for CoT or ToT as described in section 3.2

Table 2: Prompting templates for different methods in domain adaptation translation tasks. The table illustrates three distinct approaches: Standard (basic instructions), Domain-Specific (explicit domain indication in the system prompt), and Self-Guided CoT/ToT (autonomous domain inference with reasoning).

$$D = f_{\text{analyze}}(x) \tag{3}$$

$$y = f_{\text{translate}}(x, D) \tag{4}$$

where $f_{\text{analyze}} : X \to \mathcal{D}$ is a domain inference function that maps input x to domain attributes $D \in \mathcal{D}$, and $f_{\text{translate}} : X \times \mathcal{D} \to Y$ is a domain-aware translation function.

Table 2 presents three distinct prompting approaches. The Standard Prompt represents the baseline with no domain awareness. The Domain-Specific Prompt (DSP) explicitly provides domain D (Zhang et al., 2023; Vilar et al., 2023). In contrast, the Self-Guided CoT/ToT Prompt induces the model to infer D through autonomous analysis (Zhou et al., 2023; Xie et al., 2023). We evaluate these approaches across multiple domains and language pairs to assess their impact on translation quality and domain adaptation capabilities.

3.4 Model & Hyper-parameters

We conducted experiments using commercial (GPT-4o Mini) and open-source (Qwen 2.5 70B Turbo via Together AI) models. These models represent diverse architectures and training paradigms, allowing assessment across different model families. All experiments were conducted January-March 2025 using the latest available versions.

For each translation task, we applied methods from Section 3.1 and 3.2. We used a temperature of **0.6** for all generations to balance deterministic outputs with sufficient diversity. Other generation parameters included a maximum token limit of 2048, top-p value of 0.9, and no repetition penalty. For ToT prompting, we generated 3 candidate translations per input before selecting the optimal output based on the evaluation criteria described in Section 3.2.2. All prompts were implemented using the models' APIs with consistent system messages across experiments, varying only the specific prompting technique. For the domain adaptation experiments, we ensured no domain information was leaked to the models except in the explicit Domain-Specific Prompting condition.

3.5 Dataset & Evaluation

We evaluate translation capabilities across multiple dimensions: multilingual translation using **FLORES-200**(NLLB Team et al., 2024) (English, German, Mandarin Chinese, Vietnamese); domain adaptability with **WMT 2019 Biomedical**(Bawden et al., 2019), **WMT 2019 News**(Barrault et al., 2019), and **WMT 2020 Chat**(Farajian et al., 2020) datasets; and robustness to noise using **MTNT** (Michel and Neubig). For each dataset, we randomly sample from 300 to 600 sentences for evaluation. Our assessment employs three complementary metrics: **SacreBLEU** (Post, 2018) for n-gram overlap, **COMET** (Rei et al., 2020) (using the wmt22-comet-da model) for semantic adequacy, and **ChrF** (Popović, 2015) for character-level assessment particularly beneficial for morphologically rich languages. This combination provides a comprehensive evaluation of both lexical and semantic fidelity.

4 Results & Analysis

4.1 Multilingual Translation

Building upon previous findings (Peng et al., 2023; Wei et al., 2023), our research evaluates reasoning-based prompting approaches for machine translation using 50 samples from the **FLORES-200** dataset (NLLB Team et al., 2024) across four language pairs.

Table 3 demonstrates that ToT prompting with

Table 3: Impact of Reasoning Prompting on Multilingual Translation Performance

	EN→DE		DE→EN		EN→ZH		ZH→EN	
Method	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
GPT-4o Mini								
Baseline	90.56	37.23	90.63	42.31	89.60	31.53	87.81	25.83
+ 0-shot CoT	88.08↓	31.17↓	88.96↓	38.94↓	86.37↓	18.93↓	86.19↓	20.26↓
+ 1-shot CoT	87.84↓	36.19↓	89.41↓	38.63↓	87.07↓	21.21↓	86.24↓	21.77↓
+ ToT	**91.58**↑	**43.63**↑	**91.42**↑	**45.36**↑	88.98↓	29.52↓	**88.21**↑	**26.13**↑
Qwen 2.5 Turbo								
Baseline	87.83	31.34	90.35	40.81	**90.02**	34.02	88.42	31.11
+ 0-shot CoT	88.17↑	30.87↓	89.68↓	37.52↓	88.27↓	24.04↓	87.42↓	21.24↓
+ 1-shot CoT	58.89↓	10.43↓	88.58↓	37.77↓	88.45↓	28.27↓	87.66↓	22.70↓
+ ToT	88.40↑	33.43↑	89.76↑	41.47↑	90.66↑	**34.51**↑	87.97↓	26.64↓

Note: ↑/↓ indicates improvement/deterioration compared to baseline. The baseline is the result of zero-shot prompting to LLMs. Bold values highlight the best results for each language pair and metric. CoT = Chain-of-Thought, ToT = Tree-of-Thought prompting.

GPT-4o Mini significantly outperforms the baseline for European languages (+6.4 BLEU for EN→DE, +3.05 BLEU for DE→EN), while both zero-shot and one-shot CoT approaches consistently underperform across all language pairs. Qwen 2.5 Turbo shows more varied responses, with ToT improving performance for three language pairs but one-shot CoT causing catastrophic performance collapse for EN→DE (-20.91 BLEU). These patterns highlight model-specific responses to reasoning prompts (Chen et al., 2024) and ToT's superior handling of translation's branching complexity (Xie et al., 2023).

4.2 Domain Adaptation

We assess the effectiveness of reasoning-based prompting for domain adaptation in multilingual translation. Inspired by Zhou et al. (2024), we designed self-guided prompts (shown in Table 2) that enable models to autonomously infer the domain of a given text by identifying key terminology. This differs from conventional approaches that require manual domain specification (Peng et al., 2023).

We evaluate these Self-Guided Chain-of-Thought (SG-CoT) and Tree-of-Thought (SG-ToT) methods on the **WMT 2019 Biomedical** and **WMT 2019 News** datasets, comparing against standard and domain-specific baselines. Table 4 reveals three key advantages of self-guided reasoning, with SG-ToT demonstrating the strongest performance:

- **Cross-domain flexibility**: SG-ToT improves COMET scores across domains: +1.69 for EN→ZH biomedical and +0.87 for DE→EN news translation (Garcia et al., 2023).
- **Terminology consistency**: SG-ToT excels in terminology-dense contexts, achieving +4.06 BLEU (23.11 → 27.17) for ZH→EN biomedical translation with Qwen 2.5 Turbo (Peng et al., 2023).
- **Domain-adaptive accuracy**: For biomedical content, SG-ToT consistently outperforms both baseline and domain-specific prompting, with up to +2.89 BLEU improvement for ZH→EN translation (Team et al., 2022).

Interestingly, SG-CoT shows inconsistent performance, suggesting that exploring multiple translation candidates (as in ToT) is crucial for effective self-guided domain adaptation.

4.3 Noisy Texts

Building upon (Michel and Neubig), we apply our prompting methods to translate noisy text sourced from Reddit comments, containing typos, grammatical errors, code-switching, and other informalities. LLMs are tasked with translating between English (en), French (fr), and Japanese (ja). The results in Table 5 demonstrate that our approach significantly outperforms the previous work of (Michel and Neubig) in translating noisy text, highlighting the ability of modern LLMs to maintain translation quality even in the presence of data inconsistencies (Sperber et al., 2017).

ToT prompting exhibits strong performance with GPT-4o Mini, achieving the highest scores for

System	WMT19 Biomedical				WMT19 News			
	EN→ZH		ZH→EN		EN→DE		DE→EN	
	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
GPT-4o mini								
Baseline	86.10	20.89	83.32	22.53	87.65	33.16	88.29	38.14
+ DSP	**87.03**$^{\uparrow}$	**21.50**$^{\uparrow}$	**84.48**$^{\uparrow}$	23.98$^{\uparrow}$	**88.55**$^{\uparrow}$	**34.75**$^{\uparrow}$	88.48$^{\uparrow}$	**38.78**$^{\uparrow}$
+ F-DSP	86.01$^{=}$	20.51$^{\downarrow}$	83.33$^{\downarrow}$	23.00$^{\uparrow}$	87.68$^{\uparrow}$	33.30$^{\uparrow}$	88.25$^{\downarrow}$	38.13$^{\downarrow}$
+ SG-CoT	83.83$^{\downarrow}$	18.12$^{\downarrow}$	83.52$^{\uparrow}$	**25.69**$^{\uparrow}$	85.48$^{\downarrow}$	29.58$^{\downarrow}$	86.28$^{\downarrow}$	33.00$^{\downarrow}$
+ SG-ToT	**87.79**$^{\uparrow}$	**21.74**$^{\uparrow}$	83.69$^{\uparrow}$	25.42$^{\uparrow}$	88.39$^{\uparrow}$	34.58$^{\uparrow}$	**88.86**$^{\uparrow}$	38.11$^{\downarrow}$
Qwen 2.5 Turbo								
Baseline	86.55	22.70	83.40	23.11	86.17	28.83	87.99	38.28
+ DSP	86.47$^{=}$	22.62$^{=}$	83.53$^{\uparrow}$	23.18$^{\uparrow}$	86.56$^{\uparrow}$	29.37$^{\uparrow}$	88.29$^{\uparrow}$	38.81$^{\uparrow}$
+ F-DSP	86.54$^{=}$	22.59$^{=}$	83.26$^{\downarrow}$	22.93$^{\downarrow}$	86.72$^{\uparrow}$	29.17$^{\uparrow}$	88.24$^{\uparrow}$	37.74$^{\downarrow}$
+ SG-CoT	85.64$^{\downarrow}$	21.19$^{\downarrow}$	81.41$^{\downarrow}$	25.90$^{\uparrow}$	61.48$^{\downarrow}$	8.60$^{\downarrow}$	87.28$^{\downarrow}$	34.52$^{\downarrow}$
+ SG-ToT	87.08$^{\uparrow}$	22.92$^{\uparrow}$	84.39$^{\uparrow}$	27.17$^{\uparrow}$	85.26$^{\downarrow}$	28.12$^{\downarrow}$	88.85$^{\uparrow}$	37.95$^{=}$

Table 4: Translation performance comparison on WMT 2019 Biomedical and WMT 2019 News datasets. Cell colors indicate performance relative to baseline: green = improvement (darker = stronger), red = degradation, yellow = minimal change. Symbols indicate direction: ↑ = improvement, ↓ = degradation, = = no significant change. DSP = Domain-Specific Prompting, F-DSP = False Domain-Specific Prompting, SG = Self-guided, CoT = Chain-of-Thought, ToT = Tree-of-Thought. **Bold** numbers indicate best performance per column.

fr→en (38.99) and en→ja (30.54), while zero-shot and few-shot approaches also perform well in specific language pairs. Notably, CoT prompting underperforms compared to other methods, particularly with Qwen 2.5 Turbo where performance degrades substantially (e.g., only 11.65 BLEU for fr→en). This suggests that the linear reasoning process of CoT may amplify errors when handling noisy inputs (Wang et al., 2023), while ToT's exploration of multiple translation candidates provides greater robustness (Yao et al., 2023; Xie et al., 2023). Overall, GPT-4o Mini demonstrates superior performance compared to Qwen 2.5 Turbo across all prompting methods, indicating stronger resilience to textual noise in commercial models (Ateia and Kruschwitz, 2024).

4.4 Ablation Study

Tree-of-Thought: To identify essential ToT components for translation, we systematically removed individual elements and measured performance impacts (Table 6). Using the same FLORES-200 dataset from Section 4.1 with English to German (EN→DE) translation, we found that for GPT-4o Mini, candidate branching proved most critical (-8.5% when removed), while analysis and multi-dimensional evaluation showed similar importance (approximately -4.6%). Qwen 2.5 Turbo exhibited stronger dependencies, particularly on the analysis phase (-18.6%) and branching (-14.1%), suggesting open-source models benefit substantially from structured reasoning. These findings confirm that ToT's effectiveness stems from the complementary interaction of its components, with their relative importance varying by model architecture.

System	Method	Translation Direction			
		en→fr	fr→en	en→ja	ja→en
Prior Work					
Michel & Neubig (2018)	Base	21.77	23.27	9.02	6.65
Michel & Neubig (2018)	Finetuned	29.73	30.29	12.45	9.82
Our Approach					
GPT-4o Mini	Zero-shot	**38.63**	38.84	30.37	14.70
	3-shot	26.04	39.21	18.80	**15.16**
	CoT	26.46	38.01	28.28	12.91
	ToT	36.51	**38.99**	**30.54**	14.56
Qwen 2.5	Zero-shot	34.30	34.30	23.47	10.75
	3-shot	34.26	35.16	12.98	11.49
	CoT	16.36	11.65	13.59	10.38
	ToT	32.78	20.37	24.09	11.68

Table 5: BLEU scores for noisy text translation across four language directions using LLM prompting methods, compared to Michel & Neubig (2018). GPT-4o Mini's ToT prompting excels (e.g., 38.99 for fr→en, 30.54 for en→ja), with zero-shot (38.63, en→fr) and 3-shot (15.16, ja→en) also outperforming prior finetuned models. Blue shading denotes strong (light) and top (dark) scores.

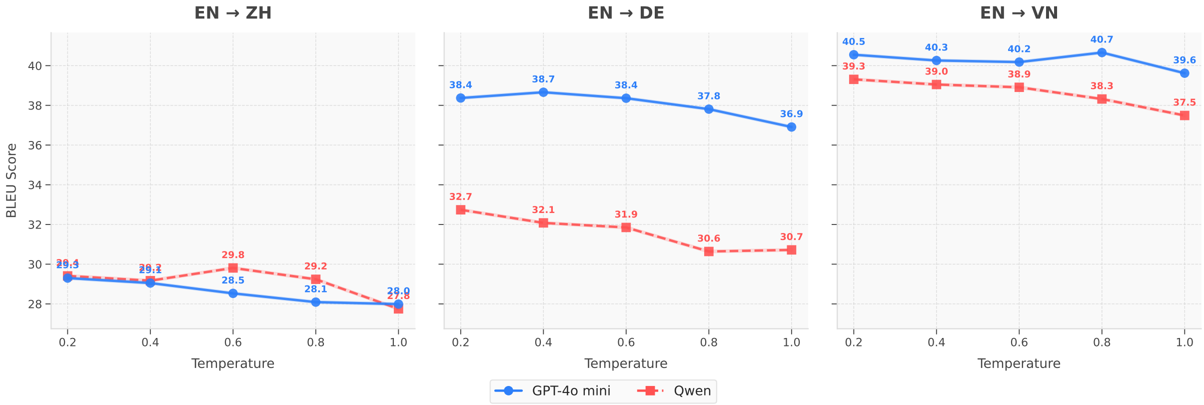

Figure 3: BLEU scores for multilingual translation across temperature settings (0.2-1.0) for English (EN) to German (DE), Chinese (ZH), and Vietnamese (VN). Higher values indicate better performance.

Table 6: Impact of ToT Components: Ablation Study Results (BLEU Scores)

Method	GPT-4o Mini		Qwen 2.5 Turbo	
	BLEU	ΔBLEU	BLEU	ΔBLEU
Full ToT (Base)	**45.26**	—	**33.43**	—
w/o Analysis	43.14	-4.7%	27.21	-18.6%
w/o Branching	41.43	-8.5%	28.70	-14.1%
w/o Multi-Evaluation	43.19	-4.6%	29.61	-11.4%
w/ Random Selection	42.35	-6.4%	33.12	-0.9%

Temperature: Temperature governs LLM text generation randomness, affecting translation faithfulness and fluency. We evaluate settings from 0.2 to 1.0 across language pairs using both lexical (BLEU) and semantic (COMET) metrics. Figures 3 and 8 reveal: (1) language-specific optimal temperatures, with EN→ZH favoring lower settings (0.2-0.4), especially for GPT-4o mini; (2) model-specific sensitivity, with GPT-4o mini showing greater performance variation across temperatures; (3) occasional BLEU and COMET trend divergence, underscoring multi-metric evaluation importance (Rei et al., 2020); and (4) performance decline at higher temperatures (near 1.0) for most language pairs. These findings highlight the necessity of language-specific temperature optimization for multilingual LLM translation (Holtzman et al., 2020).

Discussion and Future Work

Our experiments show ToT prompting significantly enhances translation accuracy for multilingual and noisy-text scenarios, outperforming CoT approaches (Yao et al., 2023). Our self-guided domain adaptation performs competitively with explicit domain-specific methods while reducing manual effort. However, these reasoning-based approaches increase computational costs, creating scalability challenges (Wu et al., 2023).

The commercial model (GPT-4o Mini) consistently outperforms the open-source alternative (Qwen 2.5 Turbo) across all prompting strategies, with this gap widening for ToT prompting. Open-source models perform adequately on simpler tasks but struggle with complex reasoning, suggesting advantages in proprietary training methodologies.

Future work includes optimizing prompt efficiency, evaluating low-resource languages (Team et al., 2022) and specialized domains, integrating prompting with fine-tuning, and conducting human-in-the-loop studies..

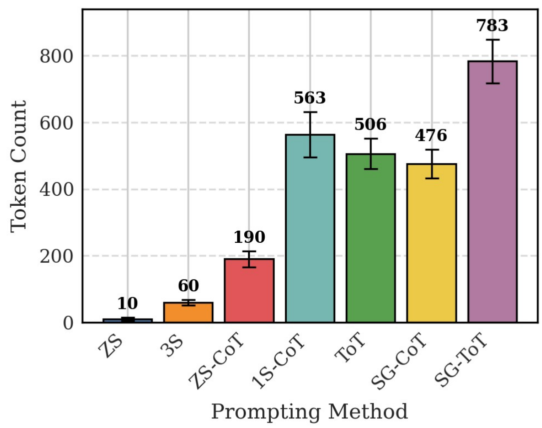

Figure 4: Token count per method. ZS = Zero-shot, 3S = Three-shot, CoT = Chain-of-Thought, ToT = Tree-of-Thought, SG = Self-guided.

Limitations

While this study provides valuable insights into reasoning-based prompting for machine translation, several limitations remain.

First, due to financial constraints, we could not evaluate a broader range of commercial and open-source models, such as **Claude 3.5 Sonnet, Llama 3.3, and Gemini 2.0 Flash**, limiting cross-architecture comparisons.

Second, **Chain-of-Thought (CoT)** and **Tree-of-Thought (ToT)** prompting incur high computational costs due to increased token usage (Figure 4), resulting in substantial API expenses (Figure 7). This may hinder accessibility, particularly for researchers with limited resources.

Finally, our experiments focus on benchmark datasets, which may not fully capture real-world domain shifts and informal text variations. Future work should explore these approaches in diverse, real-world translation scenarios to assess their robustness.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Khetam Al Sharou, Zhenhao Li, and Lucia Specia. 2021. Towards a better understanding of noise in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 53–62, Held Online. INCOMA Ltd.

Samy Ateia and Udo Kruschwitz. 2024. Can open-source llms compete with commercial models? exploring the few-shot performance of current gpt models in biomedical tasks. *Preprint*, arXiv:2407.13511.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate. *Preprint*, arXiv:1409.0473.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable 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 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. 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.

Quinten Bolding, Baohao Liao, Brandon James Denis, Jun Luo, and Christof Monz. 2023. Ask language model to clean your noisy translation data. *Preprint*, arXiv:2310.13469.

Vicent Briva-Iglesias, Joao Lucas Cavalheiro Camargo, and Gokhan Dogru. 2024. Large language models "ad referendum": How good are they at machine translation in the legal domain? *Preprint*, arXiv:2402.07681.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2024. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *Preprint*, arXiv:2310.14735.

Guanhua Chen, Shuming Ma, Yun Chen, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2022. Towards making the most of cross-lingual transfer for zero-shot neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 142–157, Dublin, Ireland. Association for Computational Linguistics.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jared Coleman, Bhaskar Krishnamachari, Ruben Rosales, and Khalil Iskarous. 2024. LLM-assisted rule based machine translation for low/no-resource languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 67–87, Mexico City, Mexico. Association for Computational Linguistics.

Maxim Enis and Mark Hopkins. 2024. From llm to nmt: Advancing low-resource machine translation with claude. *Preprint*, arXiv:2404.13813.

M. Amin Farajian, António V. Lopes, André F. T. Martins, Sameen Maruf, and Gholamreza Haffari. 2020. Findings of the WMT 2020 shared task on chat translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 65–75, Online. Association for Computational Linguistics.

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.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. *Preprint*, arXiv:2302.01398.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *Preprint*, arXiv:2302.07856.

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, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Christian Herold, Jan Rosendahl, Joris Vanvinckenroye, and Hermann Ney. 2022. Detecting various types of noise for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2542–2551, Dublin, Ireland. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models. *Preprint*, arXiv:2302.14045.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. 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, Singapore. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Preprint*, arXiv:2309.04662.

Chunyou Li, Mingtong Liu, Hongxiao Zhang, Yufeng Chen, Jinan Xu, and Ming Zhou. 2023. MT2: Towards a multi-task machine translation model with translation-specific in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8616–8627, Singapore. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *arXiv preprint arXiv:2205.05638*.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024. Chain-of-dictionary prompting elicits translation in large language models. *Preprint*, arXiv:2305.06575.

Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. 2022. Efficient machine translation domain adaptation. *Preprint*, arXiv:2204.12608.

Paul Michel and Graham Neubig. Mtnt: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *Preprint*, arXiv:2402.06196.

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

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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 EMNLP 2023*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12:26839–26874.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *Preprint*, arXiv:2009.09025.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, and 22 others. 2022. Multitask prompted training enables zero-shot task generalization. *Preprint*, arXiv:2110.08207.

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

Matthias Sperber, Jan Niehues, and Alex Waibel. 2017. Toward robust neural machine translation for noisy input sequences. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 90–96, Tokyo, Japan. International Workshop on Spoken Language Translation.

Zhixing Tan, Xiangwen Zhang, Shuo Wang, and Yang Liu. 2022. MSP: Multi-stage prompting for making pre-trained language models better translators. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6131–6142, Dublin, Ireland. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Stephan Vogel. 2003. Using noisy bilingual data for statistical machine translation.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Zhexuan Wang, Shudong Liu, Xuebo Liu, Miao Zhang, Derek Wong, and Min Zhang. 2024. Domain-aware

k-nearest-neighbor knowledge distillation for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9458–9469, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Haoze Wu, Christopher Hahn, Florian Lonsing, Makai Mann, Raghuram Ramanujan, and Clark Barrett. 2023. Lightweight online learning for sets of related problems in automated reasoning. *Preprint*, arXiv:2305.11087.

Heming Xia, Zhe Yang, Qingxiu Dong, Peiyi Wang, Yongqi Li, Tao Ge, Tianyu Liu, Wenjie Li, and Zhifang Sui. 2024. Unlocking efficiency in large language model inference: A comprehensive survey of speculative decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7655–7671, Bangkok, Thailand. Association for Computational Linguistics.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, Min-Yen Kan, Junxian He, and Qizhe Xie. 2023. Self-evaluation guided beam search for reasoning. *Preprint*, arXiv:2305.00633.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised domain adaptation for neural machine translation. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 338–343.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Bo Yuan, Yulin Chen, Yin Zhang, and Wei Jiang. 2024. Hide and seek in noise labels: Noise-robust collaborative active learning with LLMs-powered assistance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10977–11011, Bangkok, Thailand. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *Preprint*, arXiv:2301.07069.

Lei Zhang, Yuge Zhang, Kan Ren, Dongsheng Li, and Yuqing Yang. 2024. MLCopilot: Unleashing the power of large language models in solving machine learning tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2931–2959, St. Julian's, Malta. Association for Computational Linguistics.

Ran Zhang, Wei Zhao, and Steffen Eger. 2025. How good are llms for literary translation, really? literary translation evaluation with humans and llms. *Preprint*, arXiv:2410.18697.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2025. A survey of large language models. *Preprint*, arXiv:2303.18223.

Hongyi Zheng and Abulhair Saparov. 2023. Noisy exemplars make large language models more robust: A domain-agnostic behavioral analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4560–4568, Singapore. Association for Computational Linguistics.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V. Le, Ed H. Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. 2024. Self-discover: Large language models self-compose reasoning structures. *Preprint*, arXiv:2402.03620.

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 Appendix

A.1 Multilingual Translation for Zero and Few-shot Prompting

Table 7 presents results for zero-shot and few-shot translation across six language directions. Our analysis reveals language-specific strengths in the two models: GPT-4o mini excels in Germanic and Vietnamese translations with up to 7.36 BLEU points advantage for EN→DE, while Qwen 2.5 72B Turbo demonstrates superior performance in Chinese-related pairs with consistent advantages in both directions. Notably, few-shot prompting

Table 7: Zero-shot and few-shot prompting performance for multilingual translation

Model	EN→DE			EN→ZH			EN→VN		
	COMET	BLEU	ChrF	COMET	BLEU	ChrF	COMET	BLEU	ChrF
Zero-shot prompting									
GPT-4o mini	**88.78**	**38.43**	**67.33**	88.78	30.20	41.08	**89.73**	**39.45**	**60.63**
Qwen 2.5 72B Turbo	87.25	33.43	63.25	**89.02**	**30.32**	**41.05**	89.33	38.22	59.30
Few-shot prompting (3-shot)									
GPT-4o mini	**88.56**	**38.59**	**67.34**	**88.43**	29.21	40.08	**89.69**	**39.25**	**60.64**
Qwen 2.5 72B Turbo	86.15	31.23	61.72	88.18	**30.60**	**41.28**	88.67	37.72	58.60

Model	DE→EN			ZH→EN			VN→EN		
	COMET	BLEU	ChrF	COMET	BLEU	ChrF	COMET	BLEU	ChrF
Zero-shot prompting									
GPT-4o mini	**89.61**	**42.16**	**69.89**	87.32	26.77	59.74	**88.04**	**34.05**	**63.77**
Qwen 2.5 72B Turbo	89.30	40.90	69.02	**87.59**	**29.29**	**61.11**	87.01	33.67	62.90
Few-shot prompting (3-shot)									
GPT-4o mini	89.50	**41.96**	**69.72**	87.14	27.00	59.78	87.89	33.41	63.40
Qwen 2.5 72B Turbo	**89.56**	41.16	69.42	**87.25**	**27.88**	**60.53**	**87.65**	**34.35**	**64.05**

Note: Best results for each language pair and metric are in **bold**. COMET scores are multiplied by 100 for readability. EN stands for English, DE for German, ZH for Chinese, VN for Vietnamese.

does not consistently improve over zero-shot performance, contradicting patterns observed in other NLP tasks (Brown et al., 2020; Wei et al., 2022). This suggests both models possess robust internal cross-lingual representations that sufficiently handle translation without explicit examples (Johnson et al., 2017). Additionally, both models generally perform better when translating into English rather than from English, aligning with established patterns in machine translation research (Freitag et al., 2021).

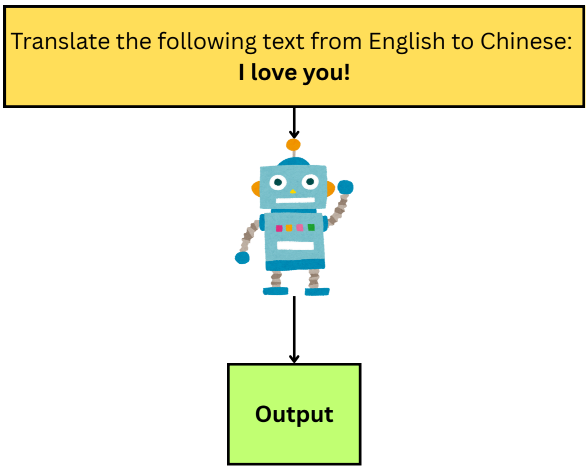

Figure 5: The workflow of zero-shot prompting

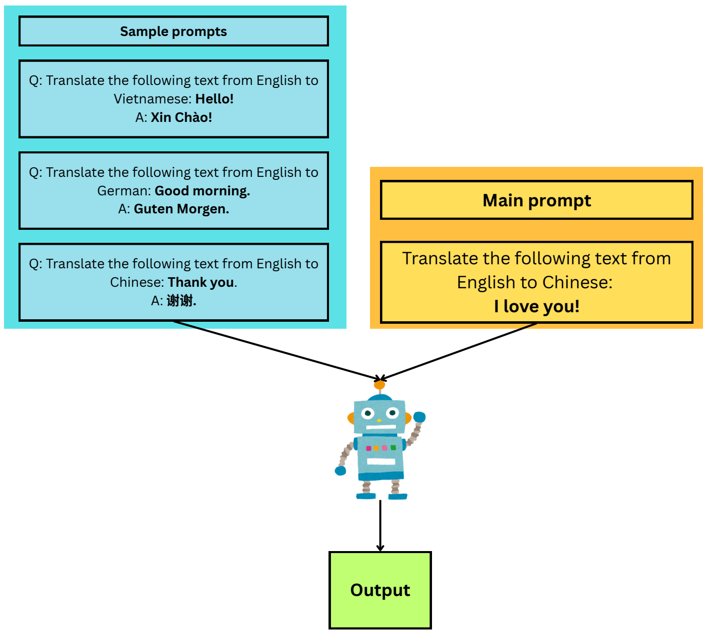

Figure 6: The workflow of few-shot prompting

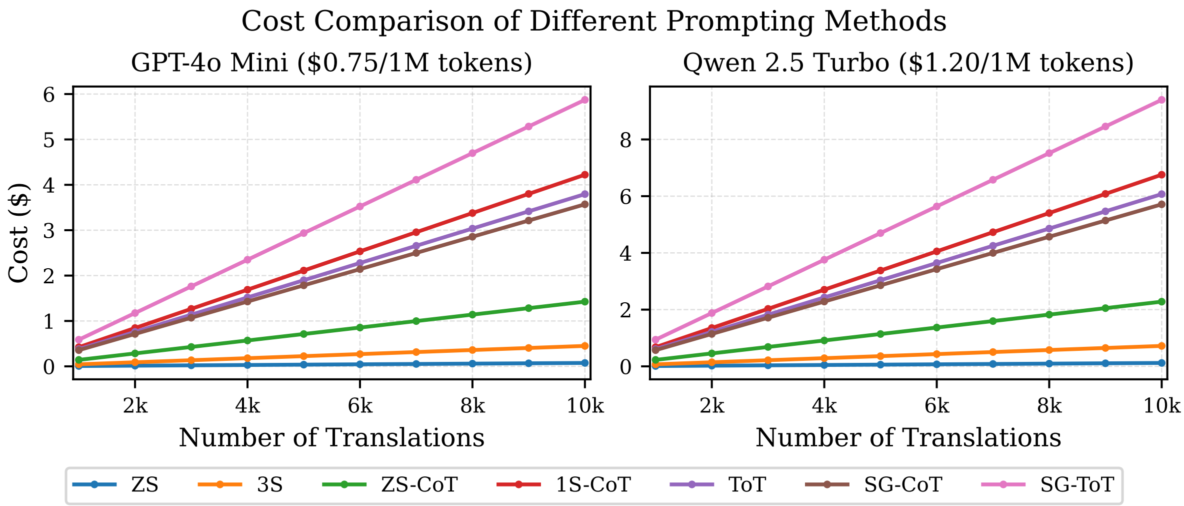

Figure 7: Cost for API calls for translation across different methods, highlighting the higher cost of reasoning prompts due to their increased token usage (Figure 4).

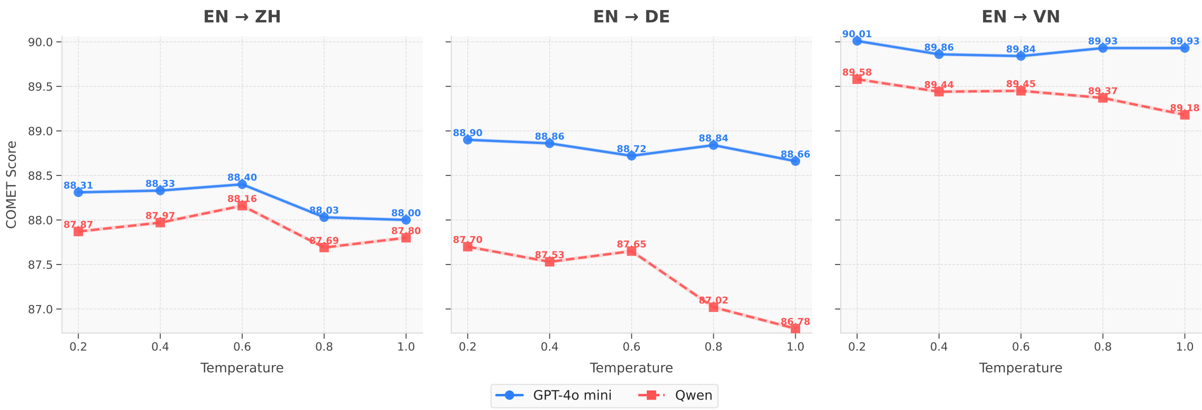

Figure 8: COMET scores for multilingual translation across temperature settings (0.2-1.0) for English (EN) to German (DE), Chinese (ZH), and Vietnamese (VN). Higher values indicate better performance.

Full Source Text	**The researchers observed crystals formed in cat urine by the addition of melamine and cyanuric acid**
Approach	**Process**
Chain-of-Thought	Step 1: The researchers observed → Die Forscher beobachteten Step 2: crystals formed → gebildete Kristalle Step 3: in cat urine → in Katzenurin Step 4: by the addition of → durch die Zugabe von Step 5: melamine → Melamin Step 6: and → und Step 7: cyanuric acid → Zyanursäure Final: Die Forscher beobachteten gebildete Kristalle in Katzenurin durch die Zugabe von Melamin und Zyanursäure.
Tree-of-Thought	Initial Analysis: • Overall Meaning: The researchers discovered crystals in cat urine that formed due to the addition of melamine and cyanuric acid. • Challenges: Technical terms (melamine, cyanuric acid), Potential ambiguity in phrasing, Scientific context may require precise translation. • Tone and Style: Formal and scientific, appropriate for academic or research contexts. Alt 1 (Literal): Die Forscher entdeckten Kristalle im Katzenurin, die durch die Zugabe von Melamin und Zyanursäure gebildet wurden. Score: 10/10 Alt 2 (Natural): Die Forscher fanden Kristalle im Urin von Katzen, die durch die Zugabe von Melamin und Zyanursäure entstanden sind. Score: 9/10 Alt 3 (Restructured): Kristalle, die im Katzenurin gefunden wurden, hatten sich durch die Zugabe von Melamin und Zyanursäure gebildet, so die Forscher. Score: 9/10 Alt 4 (Casual): Die Forscher haben Kristalle im Katzenurin entdeckt, die durch Melamin und Zyanursäure entstanden sind. Score: 8/10 Final Choice: Die Forscher entdeckten Kristalle im Katzenurin, die durch die Zugabe von Melamin und Zyanursäure gebildet wurden.

Source Text | Direct Translation | Initial Analysis | Translation Alternative | Quality Score | Final Translation

Figure 9: Translation approaches comparison (English → German) showing Chain-of-Thought and Tree-of-Thought reasoning processes.

Full Source Text	**A 28-year-old chef who had recently moved to San Francisco was found dead in the stairwell of a local mall this week**
Approach	**Process**
Chain-of-Thought	Step 1: A 28-year-old chef → Ein 28-jähriger Koch Step 2: who had recently moved to San Francisco → der kürzlich nach San Francisco gezogen war Step 3: was found dead → wurde tot aufgefunden Step 4: in the stairwell of a local mall → im Treppenhaus eines örtlichen Einkaufszentrums Step 5: this week → in dieser Woche Final: Ein 28-jähriger Koch, der kürzlich nach San Francisco gezogen war, wurde tot aufgefunden im Treppenhaus eines örtlichen Einkaufszentrums in dieser Woche.
Tree-of-Thought	Domain Analysis: News Features: journalistic style, concise reporting, factual information Requirements: clear and accessible language, maintain objectivity Alt 1 (Journalistic): Ein 28-jähriger Koch, der kürzlich nach San Francisco gezogen war, wurde diese Woche tot im Treppenhaus eines lokalen Einkaufszentrums aufgefunden. Domain Accuracy: 9/10 Technical Precision: 9/10 Overall: 9/10 Alt 2 (Balanced): Diese Woche wurde ein 28-jähriger Koch, der vor kurzem nach San Francisco gezogen ist, tot im Treppenhaus eines Einkaufszentrums gefunden. Domain Accuracy: 8/10 Audience Fit: 9/10 Overall: 8/10 Alt 3 (Accessible): Ein 28-jähriger Koch, der neu in San Francisco war, wurde diese Woche tot in einem Einkaufszentrum gefunden. Domain Accuracy: 7/10 Audience Fit: 10/10 Overall: 8/10 Final Choice: Ein 28-jähriger Koch, der kürzlich nach San Francisco gezogen war, wurde diese Woche tot im Treppenhaus eines lokalen Einkaufszentrums aufgefunden. Domain Confidence: 9/10

Source Text   Direct Translation   Domain Analysis   Translation Alternative   Evaluation Score   Final Translation

Figure 10: Domain Adaptation translation (News domain) comparison (English → German) showing Chain-of-Thought and Tree-of-Thought reasoning processes.