TACTIC: Translation Agents with Cognitive-Theoretic Interactive Collaboration

Anonymous ACL submission

Abstract

Machine translation (MT) has long been a central task in natural language processing. Large language models (LLMs) have recently delivered strong translation performance, yet fully leveraging their potential remains challenging. While multi-agent approaches show promise by decomposing translation into collaborative subtasks, they often overlook insights from Cognitive Translation Studies (CTS), which highlight how human translators flexibly combine strategies, process linguistic information, and integrate contextual knowledge. We introduce **TACTIC** (Translation Agents with Cognitive-Theoretic Interactive Collaboration), a cognitively informed multi-agent framework that operationalizes CTS principles. TACTIC comprises six specialized agents for drafting, refinement, evaluation, scoring, context reasoning, and external knowledge gathering, organized into fast translation and iterative refinement stages. This design simulates key cognitive processes of human translators and aligns translation with strategies, processing mechanisms, and contextual reasoning. On FLORES-200 and WMT24, TACTIC achieves state-of-the-art performance: using DeepSeek-V3 as the base model, it surpasses GPT-4.1 by +0.6 XCOMET and +1.18 COMETKIWI-23, and further outperforms DeepSeek-R1 by +0.84 XCOMET and +2.99 COMETKIWI-23.

1 Introduction

Machine Translation (MT) has long been modeled as conditional sequence generation, from statistical methods to neural machine translation (NMT) (Bahdanau et al., 2015; Vaswani et al., 2017). Recently, large language models (LLMs) have achieved remarkable progress in translation (Hendy et al., 2023; Lu et al., 2024), leveraging general-purpose pretraining and prompt-based conditioning. Unlike conventional NMT, LLMs translate in ways closer to human translators, integrating contextual reasoning, diverse strategies, and broad linguistic knowledge. This shift raises a fundamental question: *can we more faithfully model the cognitive processes underlying human translation?* Cognitive Translation Studies (CTS) (Halverson, 2010; Schwieter et al., 2020) offers a principled framework to address this question. CTS emphasizes the mechanisms, strategies, and contextual dimensions of translation cognition. In particular, three aspects are central: **cognitive strategies** (flexible approaches such as literal translation or free adaptation), **cognitive processing** (operations like comprehension, memory, and reformulation), and **contextual cognition** (integration of prior knowledge and discourse context).

Building on this perspective, we propose **TACTIC** (Translation Agents with Cognitive-Theoretic Interactive Collaboration)—a multi-agent framework explicitly aligned with CTS. TACTIC decomposes translation into specialized agents (*ResearchAgent*, *ContextAgent*, *DraftAgent*, *RefinementAgent*, *EvaluationAgent*, and *ScoreAgent*), organized into fast translation and iterative refinement stages. Each agent embodies specific cognitive functions, allowing the system to operationalize translation strategies, cognitive processing, and contextual reasoning in a structured manner.

Our contributions are threefold:

- **Cognitive-Inspired Multi-Agent Framework:** We propose TACTIC, a multi-agent translation architecture explicitly aligned with the core components of Cognitive Translation Studies. Each agent simulates a specific cognitive function in human translation, forming an interpretable system.

- **Operationalizing Translation Strategies and Processing:** We models strategic variation through the *DraftAgent*, which generates literal, sense-for-sense, and free translations. The *EvaluationAgent* and *ScoreAgent*

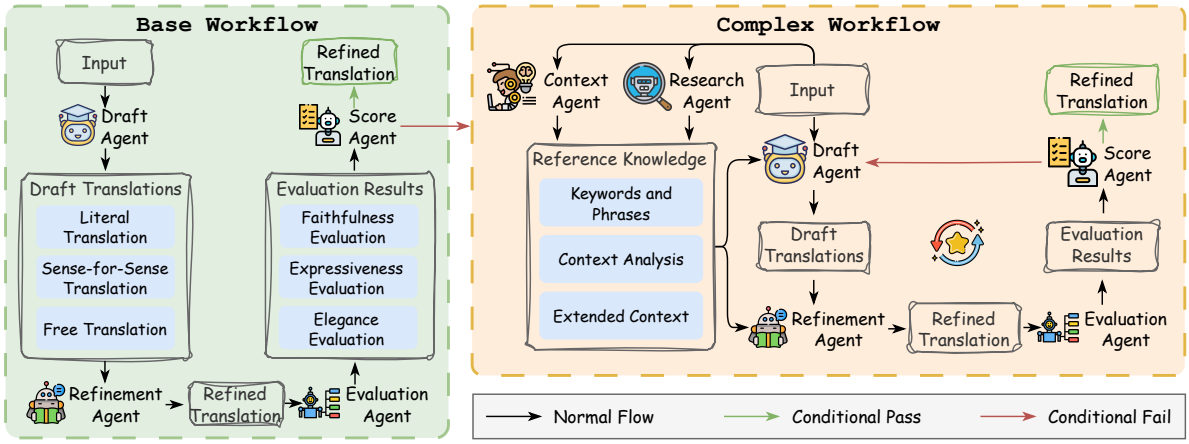

Figure 1: Overall agent collaboration workflow in the TACTIC framework. The figure depicts a cognitively inspired, modular system comprising six specialized agents: ① *DraftAgent* explores stylistic diversity through multiple translation strategies; ② *RefinementAgent* consolidates these drafts into a coherent output; ③ *EvaluationAgent* applies multidimensional quality checks; ④ *ScoreAgent* decides whether the result meets quality expectations. When challenges arise, the system escalates to a more reflective mode: ⑤ *ResearchAgent* and ⑥ *ContextAgent* inject additional knowledge and situational awareness to support iterative improvement. This dual-layered workflow reflects the flexible reasoning patterns of human translators, who shift between fluent execution and deeper analytical engagement depending on task complexity.

simulate cognitive processing by conducting feedback-based assessment and scoring grounded in the classical *faithfulness*, *expressiveness*, and *elegance* framework.

- **Contextual Cognition:** We implement the *ContextAgent* to infer and expand the discourse context of the source text, incorporating stylistic, situational, and audience-related cues to enhance contextual awareness and coherence, aligning with the contextual cognition dimension of CTS.

Together, these advances bridge human translation theory and computational practice, offering both empirical gains and theoretical insights.

2 TACTIC

Inspired by CTS, we introduce TACTIC, a modular agent-based framework designed to enhance translation quality through cognitively informed collaboration among six specialized agents. An overview of the framework is illustrated in Figure 1. Specifically, ① the *DraftAgent* applies cognitive translation strategies to generate multi-style translations, including literal, sense-for-sense, and free renditions; ② the *RefinementAgent* synthesizes these multiple drafts to produce a refined translation; ③ the *EvaluationAgent* assesses translations based on the cognitive dimensions of faithfulness, expressiveness, and elegance; ④ the *ScoreAgent* assigns a quantitative score to the refined translation based on evaluation results; ⑤ the *ResearchAgent* identifies related keywords and phrases informed by cognitive contextualization theories; and ⑥ the *ContextAgent* supplements the translation process with broader contextual information to guide and enhance agent collaboration.

The collaboration among these agents is structured into a coherent workflow. Initially, the system operates under a *base workflow*: *DraftAgent* generates multiple translation styles, *RefinementAgent* integrates them into a single refined translation, *EvaluationAgent* provides a multidimensional evaluation, and *ScoreAgent* determines whether the translation meets a pre-defined quality threshold. If the threshold is achieved, the refined translation is accepted as the final output. However, if the translation quality falls short, we are motivated to introduce a *complex workflow*, wherein *ResearchAgent* and *ContextAgent* are activated to gather domain-relevant lexical and contextual resources. These resources are then fed back into the drafting and refining stages, enabling iterative improvement until the translation meets the desired standards.

2.1 Agent Design

The design of each agent in the TACTIC framework is grounded in core principles from cognitive translation studies (Halverson, 2010) , including cogni-

CTS Concepts	TACTIC Agents
Contextual Cognition	ResearchAgent
Contextual Cognition	ContextAgent
Cognitive Strategies	DraftAgent
Cognitive Processing	RefinementAgent
Cognitive Processing	EvaluationAgent
Cognitive Processing	ScoreAgent

Table 1: Mapping Between CTS Concepts and Agents.

tive translation strategies (Kussmaul, 1995), cognitive processing models (Gile, 2009), and context-based cognitive theories (Risku, 2010), as illustrated in Table 1. These theoretical underpinnings guide the functional division and interaction protocols among agents, enabling a cognitively informed, modular architecture that mirrors human translation behavior. We introduce the detailed design of each agent as follows:

- ***DraftAgent.*** The *DraftAgent* is inspired by the cognitive strategy theory in translation, particularly the notion that translators often engage in multiple stylistic pathways (e.g., literal, sense-for-sense, and free translation) depending on task demands and textual properties. Accordingly, this agent employs a multi-style generation mechanism to produce three distinct translation variants. These styles aim to simulate the cognitive process of divergent thinking during the initial translation phase, ensuring broad semantic and stylistic coverage.

- ***RefinementAgent.*** The *RefinementAgent* synthesizes the drafts from *DraftAgent* into a single translation that is more coherent and polished. Rather than selecting the best candidate, it draws on complementary strengths across drafts to improve semantic alignment and stylistic fluency.

- ***EvaluationAgent.*** The *EvaluationAgent* is grounded in cognitive models of translation processing, with particular emphasis on internal quality control and metacognitive monitoring as observed in expert translators. It evaluates the refined translation along three cognitively motivated dimensions: *faithfulness (semantic accuracy), expressiveness (pragmatic adequacy), and elegance (stylistic naturalness)*. These dimensions are widely recognized in translation studies as essential to capturing both propositional meaning and communicative intent, as well as maintaining stylistic coherence and fluency.

- ***ScoreAgent.*** Acting as a decision-making filter, the *ScoreAgent* transforms qualitative assessments into a quantitative score that determines whether the refined translation meets a pre-defined threshold. This agent simulates the cognitive operation of performance monitoring, commonly discussed in psycholinguistic models of language production.

- ***ContextAgent.*** The *ContextAgent* is based on the theory of contextual cognition in translation, which emphasizes the role of extralinguistic knowledge and situation models. This agent enriches the translation process with relevant contextual information, including domain knowledge, register, discourse intent, the preceding and following context segments expanded by the LLMs and others, all of which are critical for ensuring pragmatic adequacy and stylistic consistency in complex translation tasks.

- ***ResearchAgent.*** Complementing the *ContextAgent*, the *ResearchAgent* focuses on lexical and conceptual elaboration. Informed by the contextual enrichment theory, this agent identifies relevant keywords, collocations, and conceptual associations that may support more informed and accurate translation choices. It functions as a dynamic external memory module, helping simulate the translator's reference-gathering behavior.

2.2 Agent Workflow

The agent workflow in TACTIC is designed not merely as a linear processing pipeline, but as a cognitively inspired collaboration protocol that mirrors how human translators dynamically allocate effort and adapt strategies based on task complexity. As illustrated in Figure 1 and formalized in Algorithm 1, the workflow unfolds in two distinct layers: a *Base Workflow* for routine cases, and a contextually enhanced *Complex Workflow* for cognitively demanding translation tasks. Let x denote the input source text and τ a predefined quality threshold. The final output T^* is produced through iterative reasoning and quality control across agents, each functioning as a specialized cognitive module.

2.2.1 Base Workflow

In the base workflow, translation begins with stylistic divergence. The *DraftAgent* generates three translation drafts, denoted as T_1, T_2, and T_3, which correspond to distinct cognitive strategies observed

Algorithm 1 TACTIC Translation Workflow

Input: source text x, quality threshold τ
Output: refined translation T^*
$\{T_1, T_2, T_3\} \leftarrow \text{DraftAgent}(x)$
$T_r \leftarrow \text{RefinementAgent}(\{T_1, T_2, T_3\})$
$(f, e, a) \leftarrow \text{EvaluationAgent}(T_r)$
$s \leftarrow \text{ScoreAgent}(f, e, a)$
if $s \geq \tau$ **then**
  $T^* \leftarrow T_r$
else
  repeat
    $K \leftarrow \text{ResearchAgent}(x)$
    $C \leftarrow \text{ContextAgent}(x)$
    $\{T'_1, T'_2, T'_3\} \leftarrow \text{DraftAgent}(x; K, C)$
    $T'_r \leftarrow \text{RefinementAgent}(\{T'_1, T'_2, T'_3\})$
    $(f', e', a') \leftarrow \text{EvaluationAgent}(T'_r)$
    $s' \leftarrow \text{ScoreAgent}(f', e', a')$
  until $s' \geq \tau$
  $T^* \leftarrow T'_r$
end if
return T^*

in human translators: *literal translation*, *sense-for-sense rendering*, and *free adaptation*. These stylistic variants emulate the divergent thinking phase in translation cognition, where multiple interpretive pathways are explored in parallel before convergence. The *RefinementAgent* then performs convergence by synthesizing these drafts into a refined candidate T_r, guided by internal heuristics for semantic cohesion and stylistic harmony. This output undergoes a triadic evaluation via the *EvaluationAgent*, which scores the translation along three core axes rooted in Chinese and cognitive translation theory: faithfulness (f), expressiveness (e), and elegance (a). The composite score $s = \text{ScoreAgent}(f, e, a)$ quantifies the overall acceptability of the translation. If $s \geq \tau$, the workflow terminates with $T^* = T_r$.

This base process is designed to be both efficient and sufficient for straightforward translation tasks, particularly in cases where the source meaning is explicit, syntactic complexity is low, and stylistic or pragmatic demands are moderate. It provides a reliable pipeline for generating high-quality translations without the need for external context or domain-specific adaptation, making it well-suited for general-purpose or low-stakes scenarios.

2.2.2 Complex Workflow

When the base workflow fails to meet the predefined quality threshold ($s < \tau$), TACTIC transitions into the complex workflow, which simulates an advanced cognitive loop inspired by expert-level translation behavior. This design reflects the notion that professional translators do not simply refine surface-level variations, but instead actively seek new knowledge, reassess the communicative context, and explore alternative interpretations when initial solutions prove inadequate.

At each iteration of this process, the system invokes two auxiliary agents: the *ResearchAgent* and the *ContextAgent*. The *ResearchAgent* extracts relevant domain-specific keywords, technical terms, idiomatic expressions, and collocations (K) that may enhance lexical precision and terminological adequacy. Simultaneously, the *ContextAgent* retrieves contextual parameters (C) that characterize the communicative situation, including but not limited to the speaker's intent, discourse framing, target audience, register, the preceding and following context segments expanded by the LLMs. Together, these agents provide an enriched cognitive grounding that supports deeper semantic interpretation and stylistic alignment.

Using the dynamically updated (K, C) pair, the *DraftAgent* generates a new triad of translation candidates $\{T'_1, T'_2, T'_3\}$, each representing a different cognitive translation strategy adapted to the enriched context. These are then synthesized by the *RefinementAgent* into a unified translation T'_r, which is subsequently assessed along the three core dimensions of faithfulness (f'), expressiveness (e'), and elegance (a') by the *EvaluationAgent*. The composite quality score $s' = \text{ScoreAgent}(f', e', a')$ determines whether the iteration has reached the quality threshold. If $s' \geq \tau$, the workflow terminates with $T^* = T'_r$ as the final translation. Otherwise, the system enters another iteration, during which the *ResearchAgent* and *ContextAgent* are re-invoked to update (K, C) based on the latest interpretive needs. The recalibration loop terminates once the score threshold or stopping conditions are satisfied.

This iterative, cognitively augmented loop allows TACTIC to engage in a form of reflective problem-solving, dynamically adjusting its interpretive scope and stylistic output based on both internal evaluation and additional informational feedback. It mirrors the human expert translator's ca-

pacity for adaptive reasoning and context-sensitive decision-making, particularly in situations involving domain complexity, pragmatic ambiguity, or high stylistic demands.

3 Experiments

3.1 Experimental Settings

Models. We employ a set of mainstream models as our backend models and baselines. The non-reasoning models include various sizes of Qwen2.5 series (Yang et al., 2024)(7B, 14B, 32B, 72B), Deepseek-V3-0324 (Liu et al., 2024), and OpenAI's GPT-4.1 (OpenAI, 2025). The reasoning-capable models include QwQ-32B (Team, 2024) and Deepseek-R1-0120 (Guo et al., 2025a). For non-reasoning models, we evaluate under both zero-shot and few-shot settings (using five examples sourced from (Xu et al., 2024a), these examples are also utilized in our non-reasoning *DraftAgent* and *RefinementAgent*). For reasoning-capable models, we adopt only the zero-shot setting.

Datasets. We evaluate our framework using two benchmark datasets: FLORES-200 (Costa-Jussà et al., 2022) and WMT24 (Kocmi et al., 2024). Our focus is on English-centric translation tasks, encompassing both English-to-X (en→xx) and X-to-English (xx→en) directions. For each dataset, we assess five language pairs: German (de), Japanese (ja), Russian (ru), Ukrainian (uk), and Chinese (zh). Since the WMT24 dataset provide only the en→xx direction, we supplement the evaluation by translating from these languages into English to analyze performance in the reverse direction. All datasets are standardized using the Tower (Alves et al., 2024) framework.

Evaluation Metrics. We focus on three primary metrics: XCOMET-XXL (Guerreiro et al., 2024) and MetricX-24-XXL (Juraska et al., 2024), both reference-based, and COMETKIWI-23-XXL (Rei et al., 2023), which is reference-free. These metrics have demonstrated high correlation with human judgments. Notably, XCOMET and MetricX were also employed in the WMT24 automatic evaluations, and we adopt their XXL variants to further enhance alignment with human assessments. In addition to these model-based metrics, we also report results on ChrF (Popović, 2015) and sacreBLEU (Post, 2018) to provide a comprehensive comparison. Due to space limitations, detailed results for these supplementary metrics are presented in Appendix A, Table 4.

3.2 Main Results

Table 2 presents the translation performance of our proposed TACTIC methods compared to baselines in both zero-shot and few-shot settings. All evaluations were conducted rigorously using both reference-based (XCOMET) and reference-free (COMETKIWI-23) metrics to ensure comprehensive and robust conclusions. Our results demonstrate that TACTIC consistently yields substantial improvements when integrated with the same underlying translation models and agent configurations. For instance, in the few-shot setting with DeepSeek-V3, TACTIC elevates the XCOMET score from 94.90 to 96.19 and improves the COMETKIWI-23 score from 90.25 to 92.64.

Our TACTIC framework, particularly when utilizing powerful models for both the agent and translation tasks, achieves state-of-the-art or highly competitive results. The configuration employing DeepSeek-V3 for both roles (TACTIC, DeepSeek-V3 // DeepSeek-V3) registers the highest XCOMET scores on FLORES-200 en→xx (**96.19**), FLORES-200 xx→en (**96.69**), and WMT24 xx→en (**89.07**). It also achieves a strong COMETKIWI-23 of 90.15 for FLORES-200 xx→en. These results often surpass strong zero-shot baselines like GPT-4.1 (FLORES-200 xx→en: XCOMET 96.50, COMETKIWI-23 88.61) and DeepSeek-R1 (WMT24 xx→en: XCOMET 87.98, COMETKIWI-23 75.01) in several categories, particularly in XCOMET scores.

The results also indicate a general trend of improved performance with larger model sizes within the same family (e.g., Qwen2.5-Instruct series) for both baseline and TACTIC settings, although TACTIC consistently provides an additional uplift. The improvements brought by TACTIC are evident across both FLORES-200 and the more challenging WMT24 dataset, and in both en→xx and xx→en translation directions, showing the robustness and broad applicability of our approach.

3.3 Ablation Studies

To assess the effectiveness of each cognitively inspired agent in TACTIC, we conducted a set of ablation studies as shown in Table 3. We can observe a clear pattern of improvement as cognitively inspired modules are added. Starting from a Zero-Shot baseline, the inclusion of the *EvaluationAgent* and *ScoreAgent* under the Iterative Evaluation setup results in noticeable quality gains. This

Table 2: Evaluation results of XCOMET and COMETKIWI-23 (abbreviated as KIWI-23) scores across different models on the FLORES-200 and WMT24 test sets. *Models* are employed for non-translation tasks (e.g., evaluation and score), whereas *Trans Models* are utilized for translation tasks (draft and refinement). All Qwen2.5 series models are Instruct versions and the "Instruct" suffix is omitted in the table for brevity. The best scores are highlighted in bold, and the second-best scores are underlined.

Models	Trans Models	FLORES-200				WMT24			
		en→xx		xx→en		en→xx		xx→en	
		XCOMET	KIWI-23	XCOMET	KIWI-23	XCOMET	KIWI-23	XCOMET	KIWI-23
Zero-shot									
Non-reasoning models									
Qwen2.5-7B	Qwen2.5-7B	77.53	70.18	91.55	85.12	66.97	59.97	81.21	77.11
Qwen2.5-14B	Qwen2.5-14B	84.17	78.25	94.05	87.17	73.97	68.57	82.59	77.83
Qwen2.5-32B	Qwen2.5-32B	86.97	81.27	94.43	87.43	76.21	70.87	84.37	79.40
Qwen2.5-72B	Qwen2.5-72B	91.45	86.14	95.71	88.20	81.23	75.81	86.00	80.48
GPT-4.1	GPT-4.1	95.46	90.83	96.50	88.61	86.83	81.03	87.73	81.33
DeepSeek-V3	DeepSeek-V3	94.43	89.80	96.14	88.45	84.71	79.61	87.20	80.99
Reasoning-capable models									
QwQ-32B	QwQ-32B	90.87	86.33	95.02	87.82	79.92	75.42	83.59	78.74
DeepSeek-R1	DeepSeek-R1	95.33	90.75	96.13	88.36	86.10	80.45	87.98	75.01
Few-shot									
Qwen2.5-7B	Qwen2.5-7B	77.50	69.73	91.06	85.00	65.25	59.03	78.44	75.09
Qwen2.5-14B	Qwen2.5-14B	85.44	79.63	94.09	87.21	75.03	69.62	82.66	77.93
Qwen2.5-32B	Qwen2.5-32B	87.65	81.87	94.31	87.28	76.61	71.22	84.36	79.48
Qwen2.5-72B	Qwen2.5-72B	91.80	86.34	95.70	88.22	81.32	76.04	85.89	80.42
GPT-4.1	GPT-4.1	95.59	91.06	96.45	88.60	86.92	81.07	87.90	81.39
DeepSeek-V3	DeepSeek-V3	94.90	90.25	96.24	88.50	85.58	80.09	87.05	81.04
TACTIC									
Non-reasoning models									
Qwen2.5-7B	Qwen2.5-7B	87.69	82.01	94.01	88.44	73.13	68.15	83.66	78.07
Qwen2.5-14B	Qwen2.5-14B	91.33	86.68	95.22	89.39	81.12	76.25	86.10	80.74
Qwen2.5-32B	Qwen2.5-32B	93.20	89.17	95.86	89.80	82.97	78.06	87.13	81.56
Qwen2.5-72B	Qwen2.5-72B	93.48	88.72	95.82	88.26	82.71	78.08	86.17	80.35
DeepSeek-V3	DeepSeek-V3	**96.19**	**92.64**	**96.69**	**90.15**	86.95	81.26	**89.07**	**82.46**
Reasoning-capable models									
Qwen2.5-7B	QwQ-32B	92.56	88.22	95.37	88.00	81.68	77.38	85.85	80.20
Qwen2.5-14B	QwQ-32B	92.90	88.51	95.70	88.19	83.47	78.65	87.20	81.54
Qwen2.5-32B	QwQ-32B	93.11	88.69	95.65	88.17	82.69	77.83	86.34	80.54
Qwen2.5-72B	QwQ-32B	93.38	88.97	95.72	88.18	82.95	78.34	86.36	80.52
Qwen2.5-32B	DeepSeek-R1	95.95	91.72	96.10	88.01	**87.60**	**81.78**	87.88	81.19
DeepSeek-V3	DeepSeek-R1	95.78	91.32	96.06	88.08	87.10	81.08	87.62	80.81

Method	XCOMET	KIWI-23
en-xx		
Zero-Shot	93.30	88.02
+ Iterative Evaluation	94.35	89.43
Few-Shot	93.45	88.17
+ Iterative Evaluation	94.37	89.29
Drafting-then-Refining	94.32	88.96
+ Iterative Evaluation	94.42	89.16
++ Keyword, Phrase and Context Mining	94.53	89.27
xx-en		
Zero-Shot	94.00	86.94
+ Iterative Evaluation	94.01	86.73
Few-Shot	94.05	86.95
+ Iterative Evaluation	94.18	86.72
Drafting-then-Refining	94.24	87.11
+ Iterative Evaluation	94.27	87.11
++ Keyword, Phrase and Context Mining	94.40	87.06

Table 3: Incremental evaluation of agent components in TACTIC. Each setting adds cognitively inspired modules, revealing steady improvements across metrics.

supports the view that translation benefits from structured self-assessment, a hallmark of expert human translators. Further improvements appear when the *DraftAgent* and *RefinementAgent* are introduced through a Drafting-then-Refining process. This reflects the widely studied two-phase model of human translation: an initial draft generation followed by semantic and stylistic adjustment. Finally, the integration of the *ResearchAgent* and *ContextAgent*, which provide keyword extraction, phrase-level cues, and discourse-level context, leads to the best observed performance. These enhancements simulate how human translators activate external knowledge and contextual frames when handling complex input, especially in low-resource or ambiguous scenarios. The results validate our design: each agent meaningfully contributes to translation quality in ways that align with cognitive theory.

4 Analysis

Synergistic Effects of TACTIC and Reasoning-Capable Models. We compare two experimental settings: (1) a TACTIC-based configuration employing DeepSeek-V3 and DeepSeek-R1; and (2) a zero-shot baseline using DeepSeek-R1 for both roles. TACTIC achieves higher scores on both XCOMET (95.78 vs. 95.33) and COMETKIWI-23 (91.32 vs. 90.75), demonstrating clear advantages. This improvement stems from TACTIC's structured, agent-based workflow, which enhances

Source:	积极运用支农支小再贷款、再贴现，引导地方法人金融机构扩大对涉农、小微和民营企业的信贷投放。
Reference:	Central bank lending for rural development and for MSBs, as well as central bank discounts, were utilized to guide locally incorporated financial institutions to expand their credit supply for agriculture-related businesses, MSBs, and private enterprises.
Zero-shot:	Actively use *rediscounts* and *relending* for agriculture and small businesses to guide *local legal person financial institutions* to expand *credit extension* to agriculture-related, small and micro, and private enterprises.
	TACTIC
Base Workflow:	Actively use *rediscounting* and *relending* for agriculture and small businesses to guide *local financial institutions* to expand credit allocation to agriculture-related, small and micro enterprises, and private enterprises.
Iteration 1:	Actively utilize relending for agriculture and small businesses and rediscounting to guide local legal entity financial institutions to expand credit allocation to agriculture-related, *micro and small enterprises*, and private enterprises.
Iteration 2:	Actively utilize *rediscount* loans for agriculture and small businesses, as well as *relending*, to guide local legal entity financial institutions to expand credit allocation to agriculture-related, small and micro, and private enterprises.
Iteration 3:	Actively utilize relending for agriculture and small businesses, as well as rediscounting, to guide local legal entity financial institutions to expand credit allocation to agriculture-related, small and micro, and private enterprises.

Figure 2: Case study demonstrating translation refinement through iterative agent collaboration. The backend model is Qwen2.5-32B-Instruct.

the deployment of reasoning capabilities by assigning focused sub-tasks to each agent. The collaborative process encourages deeper understanding of source content, enabling better fidelity, fluency, and elegance in translation.

Adaptability of Reasoning-Capable Models Within TACTIC. To isolate the impact of reasoning ability and framework structure, we compare four settings varying in model type (DeepSeek-R1 vs. DeepSeek-V3) and translation approach (zero-shot vs. TACTIC). In zero-shot settings, the reasoning-capable R1 outperforms V3. However, under TACTIC, V3 shows the greatest performance gain (XCOMET: 96.19), surpassing even the reasoning-enhanced TACTIC setup. This suggests that the benefits of structured collaboration outweigh intrinsic reasoning capabilities, especially for weaker models. TACTIC thus acts as a compensatory mechanism, amplifying translation quality by simulating decision-making processes.

Effectiveness of Iterative Refinement in TACTIC: A Case Study. We examine a case from a central bank monetary policy report in Figure 2, characterized by terminology and syntactic ambiguity. The baseline system struggles due to domain complexity. In contrast, TACTIC leverages modular agents to identify and resolve terminology, context, and style mismatches through structured refinement. Each iteration incrementally improves fidelity and fluency, ultimately producing a translation aligned with domain expectations. This validates the design of TACTIC as an effective mechanism for complex, formal-domain translation.

Distributional Impact of Iterative Refinement. To assess refinement effects, we visualize the score distribution before and after TACTIC's iterative process as shown in Figure 3. Most samples show noticeable gains, while a small subset exhibits slight regression, typically due to over-correction. The overall shift toward higher scores confirms the robustness of the agent-guided refinement mechanism, consistently improving alignment with source intent and target language norms.

Divergence Across Evaluation Dimensions. To evaluate sensitivity to nuanced translation challenges, we analyze an idiomatic example in Appendix Figure 4: "That idea went out the window." Literal, interpretive, and free strategies produce divergent outputs. RefinementAgent correctly selects the most idiomatic and contextually appropriate version. This case highlights that strategy diversity provides valuable inputs for downstream refinement, especially in figurative or culturally loaded contexts. EvaluationAgent assigns perfect scores (10/10/10) across faithfulness, expressiveness, and elegance—demonstrating its ability to discern subtle quality variations across multiple dimensions.

5 Related Work

The progress of large language models (LLMs) has reshaped machine translation, with LLM-based systems surpassing traditional NMT models (Deutsch et al., 2025). Research has explored both fine-tuned translation models (Xu et al., 2024b; Alves et al., 2024; Guo et al., 2025b) and reasoning-enhanced LLMs (He et al., 2025; Feng et al., 2025a; Wang

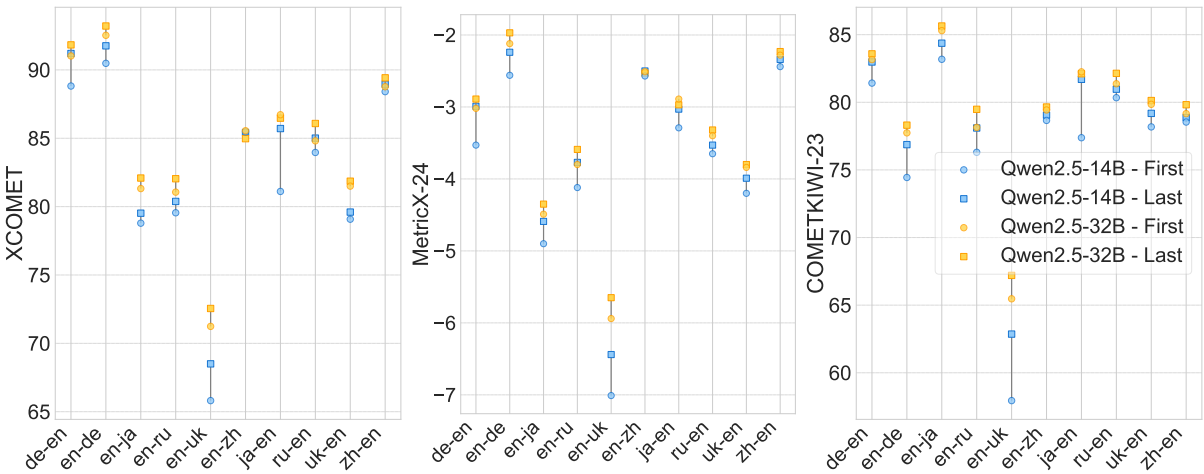

Figure 3: Distributional Impact of Iterative Refinement on the WMT24 Test Set. "Fist" and "'Last" denote the performance at the initial and final epochs, respectively.

et al., 2025b), showing that translation can benefit from domain adaptation and reasoning capabilities.

Inspired by AI agents (Li et al., 2023), multi-agent frameworks have emerged where different agents collaborate through drafting, evaluation, or refinement. Prior work includes pipeline-based designs (Briakou et al., 2024), iterative advisor–evaluator setups (Wang et al., 2025a; Feng et al., 2024), and human-like translation with keyword extraction and ranking (He et al., 2024). CRAT (Chen et al., 2024) addresses domain-specific terminology via retrieval and causal reasoning. Other systems mimic the translation publication process (Wu et al., 2024a,b), though excessive agent specialization increases complexity and hinders systematic evaluation. For translation evaluation, Feng et al. (2025b) applied MQM (The MQM Council, 2025) using multi-agent debates across four dimensions. Their framework, however, remains fragmented and computationally heavy. In contrast, we reframe evaluation into three linguistically grounded dimensions—*faithfulness*, *expressiveness*, and *elegance*—achieving a balance between interpretability and efficiency.

6 Conclusion

We propose TACTIC, a multi-agent translation framework inspired by Cognitive Translation Studies (CTS), which simulates six distinct cognitive roles: drafting, refinement, evaluation, scoring, context reasoning, and additional knowledge gathering. TACTIC supports both a base workflow and a complex workflow to accommodate varying translation scenarios. Experiments across multiple model families and evaluation metrics show that TACTIC significantly improves translation quality, achieving state-of-the-art performance on the FLORES-200 and WMT24 test sets.

Limitations

While our framework demonstrates promising progress toward cognitively inspired machine translation, several limitations remain. The multi-agent structure introduces additional complexity and computational cost, which may affect scalability and real-time performance. Moreover, our current design captures only a subset of the rich cognitive mechanisms underlying human translation, leaving room for deeper theoretical integration. Future work will explore more efficient coordination strategies, broader linguistic and domain coverage, and closer alignment with empirical findings from Cognitive Translation Studies to enhance both interpretability and generalization.

Ethics Statement

This work uses publicly available datasets and open-source models, without involving human subjects or sensitive data. TACTIC aims to enhance the interpretability of machine translation through cognitively inspired modular design, rather than replicating human cognition. While no direct ethical risks are foreseen, we acknowledge potential biases inherited from pretrained models and address them through broad evaluation and qualitative inspection. The framework is intended for research purposes, with human oversight recommended for sensitive or high-stakes translation scenarios.

References

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, and 1 others. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR*.

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. *Preprint*, arXiv:2409.06790.

Meiqi Chen, Fandong Meng, Yingxue Zhang, Yan Zhang, and Jie Zhou. 2024. Crat: A multi-agent framework for causality-enhanced reflective and retrieval-augmented translation with large language models. *Preprint*, arXiv:2410.21067.

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*.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, and 1 others. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *arXiv preprint arXiv:2502.12404*.

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

Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahan Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2025b. M-mad: Multidimensional multi-agent debate for advanced machine translation evaluation. *Preprint*, arXiv:2412.20127.

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Tear: Improving llm-based machine translation with systematic self-refinement. *Preprint*, arXiv:2402.16379.

Daniel Gile. 2009. *Basic concepts and models for interpreter and translator training*. John Benjamins Publishing Company.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hongcheng Guo, Fei Zhao, Shaosheng Cao, Xinze Lyu, Ziyan Liu, Yue Wang, Boyang Wang, Zhoujun Li, Chonggang Lu, Zhe Xu, and 1 others. 2025b. Redefining machine translation on social network services with large language models. *arXiv preprint arXiv:2504.07901*.

Sandra L Halverson. 2010. Cognitive translation studies: Developments in theory and method. In *Translation and cognition*, pages 349–369. John Benjamins Publishing Company.

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*.

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

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? A comprehensive evaluation. *CoRR*, abs/2302.09210.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. Metricx-24: The google submission to the wmt 2024 metrics shared task. *Preprint*, arXiv:2410.03983.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Paul Kussmaul. 1995. *Training the translator*. John Benjamins Publishing Company.

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

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. Llamax: Scaling linguistic horizons of LLM by enhancing translation capabilities beyond 100 languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 10748–10772. Association for Computational Linguistics.

OpenAI. 2025. Introducing gpt-4.1 in the api. Accessed: 2025-04-23.

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.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Ricardo Rei, Nuno M Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José GC de Souza, and André FT Martins. 2023. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task. *arXiv preprint arXiv:2309.11925*.

Hanna Risku. 2010. A cognitive scientific view on technical communication and translation: Do embodiment and situatedness really make a difference? *Target. International Journal of Translation Studies*, 22(1):94–111.

John W Schwieter, Julia Festman, and Aline Ferreira. 2020. Current research in bilingualism and its implications for cognitive translation and interpreting studies. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, 19.

Qwen Team. 2024. Qwq-32b: Embracing the power of reinforcement learning. Accessed: 2025-04-23.

The MQM Council. 2025. The mqm error typology. `https://themqm.org/error-types-2/typology/`. Accessed: 2025-04-27.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NeurIPS*, pages 5998–6008.

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. *arXiv preprint arXiv:2504.10187*.

Minghao Wu, Jiahao Xu, and Longyue Wang. 2024a. TransAgents: Build your translation company with language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 131–141, Miami, Florida, USA. Association for Computational Linguistics.

Minghao Wu, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. 2024b. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Preprint*, arXiv:2405.11804.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2024b. X-alma: Plug & play modules and adaptive rejection for quality translation at scale. *arXiv preprint arXiv:2410.03115*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

A Additional Results

As shown in Table 4, Lexical-based metrics such as BLEU and ChrF exhibit clear inconsistencies compared to model-based metrics like XCOMET, COMETKIWI-23, and MetricX-24. Consistent with the findings of Deutsch et al. (2025), model-based metrics demonstrate strong internal agreement, consistently assigning higher scores to leading LLMs (e.g., OpenAI o1, Claude, and Gemini), and aligning more closely with human evaluations. In contrast, lexical-based metrics often favor traditional machine translation systems such as *Google Translate* and *DeepL Translate*. Therefore, we include BLEU and ChrF results for completeness.

In addition, we provide supplementary results in Tables 5 and 6 for completeness.

Figure 4: Case study visualizing different translation strategies and evaluation dimensions. The backend model is Qwen2.5-32B-Instruct.

B Implementation Details

All open-source models, except DeepSeek-V3 and DeepSeek-R1, are primarily deployed locally using `vLLM` (Kwon et al., 2023) on machines each equipped with 8 × NVIDIA A100 GPUs (40GB per GPU). In contrast, DeepSeek-V3 and DeepSeek-R1 are deployed in a distributed manner across multiple machines, each equipped with 8 × Ascend 910B NPUs (64GB per NPU). For all models except DeepSeek-V3, the output parameters are consistently set to `max_model_len` = 8192, `max_tokens` = 4096, and `temperature` = 0.6. For DeepSeek-V3, we follow the official recommendation and adopt a lower temperature of 0.3. Due to the inherent randomness of LLMs outputs, individual experimental results may exhibit slight fluctuations, however, the overall distribution remains stable. In edge-case scenarios or when processing syntactically complex inputs, TACTIC may enter a prolonged iterative cycle. To ensure the framework remains robust across diverse conditions and produces translations within a controllable time frame, we introduce two additional constraints: a maximum iteration threshold (κ) and a maximum execution time threshold (δ). Once either threshold is reached, the framework outputs the highest-scoring translation generated during the iterative process.

C Error Analysis

We encountered several practical issues while experimenting with our framework. Below is a concise summary:

- **External API instability**: When using external APIs as backend models, we occasionally experienced request failures. To improve robustness, we implemented an automatic retry mechanism upon failure.
- **Strict JSON formatting requirements**: Given that our agent framework involves extensive data parsing, strict control over JSON-formatted outputs is necessary. For external APIs, we manually verify whether JSON output is supported via their official documentation. For local deployment, we utilize structured output support provided by *vLLM*.
- **Repetitive output in local inference**: When running inference with locally deployed models, we occasionally observed endless repetition in output. For example, on a WMT24 test sample "@user47 noooooOOOOooOoOoooooo", smaller models (e.g., 7B, 14B) may repeatedly output "@user47 noooooOOOOooOoOooo..." until the maximum token length is reached. To mitigate this, we implemented an overlength detection and retry mechanism.
- **Untranslated inputs**: For certain source texts such as "@user8", smaller models may fail to generate valid translations. In such cases, we enable a retry mechanism to ensure output quality.

D Prompts

DraftAgent

```
SYSTEM_PROMPT
You are a native speaker of both {source_language} and {
target_language}, with expertise in translating from {
source_language} to {target_language}.

USER_PROMPT
```

Table 4: Evaluation results of ChrF, BLEU, and MetricX-24 (abbreviated as MetricX) scores across different models on the FLORES-200 and WMT24 test sets. *Models* are employed for non-translation tasks (e.g., evaluation and score), whereas *Trans Models* are utilized for translation tasks (draft and refinement). All Qwen2.5 series models are Instruct versions and the "Instruct" suffix is omitted in the table for brevity. The best scores are highlighted in bold, and the second-best scores are underlined.

Models	Trans Models	FLORES-200						WMT24					
		en→xx			xx→en			en→xx			xx→en		
		ChrF	BLEU	MetricX	ChrF	BLEU	MetricX	ChrF	BLEU	MetricX	ChrF	BLEU	MetricX
Zero-shot													
Non-reasoning models													
Qwen2.5-7B	Qwen2.5-7B	42.55	25.32	-5.66	59.60	31.25	-2.64	38.59	21.44	-6.91	52.35	24.41	-4.04
Qwen2.5-14B	Qwen2.5-14B	46.24	28.85	-4.07	62.14	33.28	-2.05	42.35	24.35	-5.11	54.12	26.73	-3.76
Qwen2.5-32B	Qwen2.5-32B	48.44	31.28	-3.49	62.68	34.12	-2.00	43.43	25.58	-4.75	54.87	27.15	-3.36
Qwen2.5-72B	Qwen2.5-72B	51.04	33.89	-2.64	64.00	36.02	-1.73	46.10	28.25	-3.81	56.17	28.32	-3.02
GPT-4.1	GPT-4.1	**55.45**	39.00	-1.76	**64.99**	36.35	**-1.57**	49.65	31.51	-2.81	57.50	29.06	-2.73
DeepSeek-V3	DeepSeek-V3	55.22	38.84	-1.94	64.42	36.46	-1.68	**49.70**	**32.41**	-3.11	57.32	29.58	-2.87
Reasoning-capable models													
QwQ-32B	QwQ-32B	46.48	26.00	-2.69	61.47	30.98	-1.86	40.88	19.77	-3.99	54.54	26.57	-3.45
DeepSeek-R1	DeepSeek-R1	53.37	35.98	-1.77	63.16	34.04	-1.62	46.78	28.08	-2.88	**71.59**	**51.72**	**-2.13**
Few-shot													
Qwen2.5-7B	Qwen2.5-7B	42.35	25.20	-5.83	59.50	30.63	-2.73	38.16	20.56	-7.14	51.49	23.56	-4.40
Qwen2.5-14B	Qwen2.5-14B	46.54	29.22	-3.76	61.94	33.22	-2.03	42.28	24.16	-4.90	53.93	26.23	-3.69
Qwen2.5-32B	Qwen2.5-32B	48.34	31.53	-3.40	62.39	34.14	-2.01	43.40	25.80	-4.68	55.12	27.02	-3.27
Qwen2.5-72B	Qwen2.5-72B	51.13	34.15	-2.59	63.86	35.71	-1.71	45.97	28.11	-3.76	56.14	28.16	-3.01
GPT-4.1	GPT-4.1	55.14	38.49	-1.71	64.83	36.31	**-1.57**	49.61	31.41	-2.77	57.62	29.38	-2.73
DeepSeek-V3	DeepSeek-V3	55.26	**39.06**	-1.87	64.62	**37.05**	-1.64	49.56	32.29	-3.01	57.21	29.52	-2.85
TACTIC													
Non-reasoning models													
Qwen2.5-7B	Qwen2.5-7B	46.04	28.02	-3.55	60.13	30.35	-2.11	40.47	22.01	-5.33	52.37	22.76	-3.60
Qwen2.5-14B	Qwen2.5-14B	48.67	30.96	-2.69	61.67	32.00	-1.84	43.66	24.66	-3.91	54.08	24.86	-3.18
Qwen2.5-32B	Qwen2.5-32B	50.39	33.07	-2.41	62.40	33.29	-1.77	44.98	26.57	-3.62	54.75	25.78	-3.04
Qwen2.5-72B	Qwen2.5-72B	51.61	34.34	-2.21	63.08	34.27	-1.71	46.22	27.83	-3.41	55.42	26.92	-2.98
DeepSeek-V3	DeepSeek-V3	54.29	37.82	-1.71	63.72	35.62	-1.60	48.79	31.35	-2.77	56.78	28.82	-2.75
Reasoning-capable models													
Qwen2.5-7B	QwQ-32B	49.70	31.73	-2.32	61.28	31.27	-1.78	43.54	24.08	-3.57	53.46	23.92	-3.05
Qwen2.5-14B	QwQ-32B	49.97	31.89	-2.28	61.70	31.98	-1.72	43.88	24.51	-3.47	53.90	24.55	-2.98
Qwen2.5-32B	QwQ-32B	49.94	31.78	-2.24	61.69	31.79	-1.73	43.94	24.62	-3.48	53.80	24.41	-2.98
Qwen2.5-72B	QwQ-32B	50.23	32.11	-2.18	61.87	32.18	-1.70	44.12	24.71	-3.38	53.93	24.51	-2.95
Qwen2.5-32B	DeepSeek-R1	52.19	33.65	**-1.62**	61.75	31.96	-1.61	45.10	25.03	-2.69	53.78	24.40	-2.77
DeepSeek-V3	DeepSeek-R1	52.46	33.93	-1.64	61.91	32.22	-1.60	45.56	25.80	**-2.68**	54.39	25.14	-2.76

```
In this phase, your sole objective is to generate a draft
translation that strictly adheres to the source text. Avoid
adding any additional information not present in the source
text, nor omit any content from it. Every detail must be fully
 preserved in the {target_language} translation. Below are
several translation strategies. Please provide your best {
translation_type} for the following source text.

Translation Strategies:
1. Literal Translation: Also known as direct translation or
word-for-word translation, it prioritizes accurate meaning
while preserving the original text's form and content in the
target language.
2.Sense-for-Sense Translation: Focuses on conveying the core
meaning of the original text without adhering strictly to its
linguistic form, ensuring greater fluency and naturalness in
the target language.
3.Free Translation: Emphasizes delivering the overall meaning
and effect of the original text, allowing for significant
rewriting or restructuring as needed.

## Pre-translation Research:
{pre_translation_result}

## Context Analysis:
{context_analysis}

## Extended Context:
{extended_context}

## Few-shot Examples:
{few_shot_examples}

## Source Text ({source_language}):
{source_text}

## Output format specification:
```json
{{
    "translation": "<Your translation>"
}}
```
The JSON object: json
```

RefinementAgent

```
SYSTEM_PROMPT
You are a native speaker of both {source_language} and {
target_language}, with expertise in translating from {
source_language} to {target_language}. As an assistant
dedicated to enhancing translation quality, you will be given
a source sentence in {source_language}, a list of candidate
translations in {target_language}, along with relevant
research. Your task is to carefully analyze the provided
information and refine the translation, ensuring it accurately
 and fully captures the original meaning of the source text.
Your analysis should be in English.

USER_PROMPT
Refine the translation from {source_language} to {
target_language}.

## Pre-translation Research:
{pre_translation_result}

## Context Analysis:
{context_analysis}

## Extended Context:
{extended_context}

## Few-shot Examples:
{few_shot_examples}

## Source text ({source_language}):
{source_text}
```

Table 5: Evaluation results (XCOMET and MetricX-24, abbreviated as MetricX) for EN-ZH and ZH-EN translation across different models on the FLORES-200 and WMT24 test sets. *Models* are employed for non-translation tasks (e.g., evaluation and score), whereas *Trans Models* are utilized for translation tasks (draft and refinement). All Qwen2.5 series models are Instruct versions and the "Instruct" suffix is omitted in the table for brevity. The best scores are highlighted in bold, and the second-best scores are underlined.

Models	Trans Models	FLORES-200				WMT24			
		en→zh		zh→en		en→zh		zh→en	
		XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX
Zero-shot									
Non-reasoning models									
Qwen2.5-7B	Qwen2.5-7B	88.26	-2.76	92.89	-1.76	79.06	-3.53	85.39	-2.92
Qwen2.5-14B	Qwen2.5-14B	91.52	-2.04	95.22	-1.30	82.07	-2.83	86.03	-2.63
Qwen2.5-32B	Qwen2.5-32B	90.84	-2.12	95.42	-1.25	81.19	-3.00	86.21	-2.69
Qwen2.5-72B	Qwen2.5-72B	92.63	-1.84	96.88	-0.92	83.67	-2.61	87.92	-2.22
GPT-4.1	GPT-4.1	93.97	-1.56	97.19	**-0.82**	85.74	-2.30	89.18	-1.98
DeepSeek-V3	DeepSeek-V3	92.51	-1.80	96.92	-0.91	83.08	-2.66	88.07	-2.22
Reasoning-capable models									
QwQ-32B	QwQ-32B	93.23	-1.69	96.41	-1.00	84.23	-2.48	85.81	-2.66
DeepSeek-R1	DeepSeek-R1	94.38	-1.49	97.09	-0.84	85.94	-2.23	88.84	-2.05
Few-shot									
Qwen2.5-7B	Qwen2.5-7B	89.04	-2.66	93.70	-1.60	78.22	-3.51	83.93	-3.17
Qwen2.5-14B	Qwen2.5-14B	91.66	-1.95	95.46	-1.24	82.74	-2.70	85.40	-2.80
Qwen2.5-32B	Qwen2.5-32B	91.61	-1.92	95.63	-1.18	82.12	-2.77	86.17	-2.55
Qwen2.5-72B	Qwen2.5-72B	93.20	-1.74	97.03	-0.87	84.26	-2.52	87.33	-2.26
GPT-4.1	GPT-4.1	94.17	-1.50	97.06	**-0.82**	86.39	-2.25	89.18	-2.02
DeepSeek-V3	DeepSeek-V3	93.15	-1.68	96.81	-0.92	83.94	-2.53	87.70	-2.24
TACTIC									
Non-reasoning models									
Qwen2.5-7B	Qwen2.5-7B	91.43	-2.00	95.96	-1.09	80.39	-2.93	86.55	-2.53
Qwen2.5-14B	Qwen2.5-14B	92.50	-1.80	96.58	-0.99	83.30	-2.62	87.13	-2.38
Qwen2.5-32B	Qwen2.5-32B	93.03	-1.73	96.39	-0.98	83.37	-2.55	87.52	-2.32
Qwen2.5-72B	Qwen2.5-72B	93.09	-1.69	96.79	-0.91	83.94	-2.54	88.02	-2.23
DeepSeek-V3	DeepSeek-V3	**95.56**	-1.43	**97.66**	**-0.82**	**87.06**	**-2.07**	**90.89**	**-1.97**
Reasoning-capable models									
Qwen2.5-7B	QwQ-32B	93.51	-1.61	96.37	-0.98	84.82	-2.34	87.92	-2.19
Qwen2.5-14B	QwQ-32B	93.79	-1.60	96.88	-0.91	85.40	-2.29	88.30	-2.17
Qwen2.5-32B	QwQ-32B	93.84	-1.59	96.51	-0.93	85.34	-2.30	88.30	-2.15
Qwen2.5-72B	QwQ-32B	93.80	-1.55	96.57	-0.92	85.54	-2.29	88.42	-2.15
Qwen2.5-14B	DeepSeek-R1	94.83	-1.42	97.02	-0.85	86.56	**-2.07**	88.92	-2.06
Qwen2.5-32B	DeepSeek-R1	94.90	**-1.41**	97.00	-0.85	86.67	-2.16	88.91	-2.11
DeepSeek-V3	DeepSeek-R1	94.84	-1.44	97.04	-0.86	86.77	-2.10	88.95	-2.03

```
## Candidate translations ({target_language}):
{candidate_translations}

## Output format specification:
```json
{{
    "analysis": "<Brief analysis of the candidate translations
>",
    "translation": "<Your refined translation>"
}}
```
The JSON object: json
```

EvaluationAgent

```
SYSTEM_PROMPT
You are a native speaker of both {source_language} and {
target_language}, as well as an expert in translation. Given a
 source sentence and its translation, your task is to assess
the quality of the translation and provide suggestions for
improvement.

USER_PROMPT
Please evaluate the faithfulness, expressiveness and elegance
of the given translation based on the provided criteria:

Faithfulness Evaluation Criteria:
1.Addition: Translation includes information not present in
the source.
2.Omission: Translation is missing content from the source.
3.Mistranslation: Translation does not accurately represent
the source.
4.Untranslated text: Source text has been left untranslated.

Expressiveness Evaluation Criteria:
1.Punctuation: Incorrect punctuation (for locale or style).
2.Spelling: Incorrect spelling or capitalization.
3.Grammar: Problems with grammar, other than orthography.
4.Register: Wrong grammatical register (eg, inappropriately
informal pronouns).
5.Inconsistency: Internal inconsistency (not related to
terminology).
6.Character encoding: Characters are garbled due to incorrect
encoding.

Elegance Evaluation Criteria:
1.Terminology: Terminology is either non-standard, does not
fit the context, or is used inconsistently.
2.Style: Translation has stylistic problems.
3.Locale convention: Wrong format for addresses, currency,
dates, names, telephone numbers, time expressions, or other
locale-specific elements.
4.Logical Expression: Translation lacks logical coherence or
does not align with the thinking patterns and language
expressions of the target language.
5.Other: Any other issues that might affect the elegance of
the translation.

## Source Text ({source_language}):
{source_text}

## Translation ({target_language}):
{translation}

## Output format specification:
```json
{{
    "faithfulness": "<Your assessment based on the
faithfulness criteria.>",
    "expressiveness": "<Your assessment based on the
expressiveness criteria.>",
    "elegance": "<Your assessment based on the elegance
criteria.>"
```
```

Table 6: Comparison results of the TACTIC framework under zero-shot and few-shot (default) prompt settings, applied to the DraftAgent and RefinementAgent using the Qwen3-32B model in non-inference mode. Overall, few-shot performs better for en-zh, while zero-shot is superior for zh-en.

Prompts	Directions	Datasets	ChrF	BLEU	TER	XCOMET	COMETKIWI-23	MetricX-24
zero-shot	en-zh	FLORES-200	39.01	45.84	100.14	93.78	89.18	-1.74
		WMT24	37.05	38.81	102.09	85.51	78.47	-2.55
	zh-en	FLORES-200	59.41	28.13	61.59	96.97	88.89	-0.97
		WMT24	53.28	23.64	67.32	89.29	79.50	-2.26
few-shot	en-zh	FLORES-200	38.52	45.30	99.95	94.16	89.60	-1.70
		WMT24	36.70	38.47	98.75	85.87	79.46	-2.45
	zh-en	FLORES-200	59.14	27.65	61.92	96.86	88.79	-0.99
		WMT24	53.25	23.40	67.79	88.55	79.18	-2.38

}}
```
The JSON object: json

### *ScoreAgent*

SYSTEM_PROMPT
You are a native speaker of both {source_language} and {target_language}, as well as an expert in translation. As an assistant dedicated to enhancing translation quality, you will be provided with a source sentence in {source_language}, its translation in {target_language}, and three evaluation criteria. Your task is to assess the translation based on three factors: faithfulness (10 points), expressiveness (10 points), and elegance (10 points). Each score ranges from 1 to 10, where lower scores indicate poorer quality, and higher scores indicate better quality. A score of 10 indicates that the translation is flawless in that specific dimension.

USER_PROMPT
Please score the translation based on the following criteria, evaluations, and source text.

Faithfulness Evaluation Criteria:
1. Addition: Translation includes information not present in the source.
2. Omission: Translation is missing content from the source.
3. Mistranslation: Translation does not accurately represent the source.
4. Untranslated text: Source text has been left untranslated.

Expressiveness Evaluation Criteria:
1. Punctuation: Incorrect punctuation (for locale or style).
2. Spelling: Incorrect spelling or capitalization.
3. Grammar: Problems with grammar, other than orthography.
4. Register: Wrong grammatical register (e.g., inappropriately informal pronouns).
5. Inconsistency: Internal inconsistency (not related to terminology).
6. Character encoding: Characters are garbled due to incorrect encoding.

Elegance Evaluation Criteria:
1. Terminology: Terminology is either non-standard, does not fit the context, or is used inconsistently.
2. Style: Translation has stylistic problems.
3. Locale convention: Wrong format for addresses, currency, dates, names, telephone numbers, time expressions, or other locale-specific elements.
4. Logical Expression: Translation lacks logical coherence or does not align with the thinking patterns and language expressions of the target language.
5. Other: Any other issues that might affect the elegance of the translation.

## Source Text ({source_language}):
{source_text}

## Translation ({target_language}):
{translation}

## Evaluation:
{evaluation_result}

## Output format specification:
```json
{{
    "faithfulness_score": "<faithfulness score>",
    "expressiveness_score": "<expressiveness score>",
    "elegance_score": "<elegance score>",
    "overall_score": "<The sum of faithfulness_score, expressiveness_score and elegance_score.>",
    "feedback": "<Brief feedback for each criterion, explaining your score.>"
}}
```
The JSON object: json

ContextAgent

SYSTEM_PROMPT
You are a native speaker of both {source_language} and {target_language}, with expertise in translating from {source_language} to {target_language}.

USER_PROMPT
Below is a source sentence in {source_language}. Your task is to infer the possible context, including style, purpose, target audience and other environmental factors that maybe helpful in understanding the source text, Your context analysis should be in English. Then, expand the source sentence by adding a previous sentence and a next sentence, using the same language, together, these sentences should form a coherent and complete passage.

Source Text ({source_language}):
{source_text}

Output format specification:
```json
{{
    "context_analysis": "<Your context analysis>",
    "extended_context": "Expanded previous sentence. source sentence. Expanded next sentence."
}}
```
The JSON object: json

ResearchAgent

SYSTEM_PROMPT
You are a native speaker of both {source_language} and {target_language}, with expertise in translating from {source_language} to {target_language}.

USER_PROMPT
Before translation, conducting thorough pre-translation research is crucial to identifying elements of the text that may pose translation challenges. The objective at this stage is to compile a list of keywords and phrases essential for accurately understanding the source text. These may include technical terms, proper names, idiomatic expressions, or other context-dependent elements that require special attention to ensure precise translation.

Source Text ({source_language}):
{source_text}

Please list keywords and phrases directly in {source_language} and {target_language} using the following output format. Do not generate any other content:

1. keyword1: keyword1 in {target_language}
2. phrases2: phrases2 in {target_language}
```

Zero-shot

```
SYSTEM_PROMPT
You are a helpful assistant.
USER_PROMPT
Translate this from {source_language} to {target_language}:
{source_language}: {source_text}
{target_language}:

Please generate the final translation in JSON format as
follows:
## Output format specification:
```json
{{
    "translation": "<Your translation>"
}}
```
The JSON object: json
```

Few-shot

```
SYSTEM_PROMPT
You are a helpful assistant.
USER_PROMPT
Translate this from {source_language} to {target_language}:
{few_shot_examples}

{source_language}: {source_text}
{target_language}:

Please generate the final translation in JSON format as
follows:
## Output format specification:
```json
{{
    "translation": "<Your translation>"
}}
```
The JSON object: json
```