Test-Time Scaling of Reasoning Models for Machine Translation

Zihao Li,[1] **Shaoxiong Ji,**[2,3,1] **Jörg Tiedemann**[1]
[1]University of Helsinki [2] ELLIS Institute Finland [3] University of Turku
`firstname.lastname@{`[1]`helsinki.fi,` [3]`utu.fi}`

Abstract

Test-time scaling (TTS) has enhanced the performance of Reasoning Models (RMs) on various tasks such as math and coding, yet its efficacy in machine translation (MT) remains underexplored. This paper investigates whether increased inference-time computation improves translation quality. We evaluate 12 RMs across a diverse suite of MT benchmarks spanning multiple domains, examining three scenarios: direct translation, forced-reasoning extrapolation, and post-editing. Our findings show that for general-purpose RMs, TTS provides limited and inconsistent benefits for direct translation, with performance quickly plateauing. However, the effectiveness of TTS is unlocked by domain-specific fine-tuning, which aligns a model's reasoning process with task requirements, leading to consistent improvements up to an optimal, self-determined reasoning depth. We also find that forcing a model to reason beyond its natural stopping point consistently degrades translation quality. In contrast, TTS proves highly effective in a post-editing context, reliably turning self-correction into a beneficial process. These results indicate that the value of inference-time computation in MT lies not in enhancing single-pass translation with general models, but in targeted applications like multi-step, self-correction workflows and in conjunction with task-specialized models.

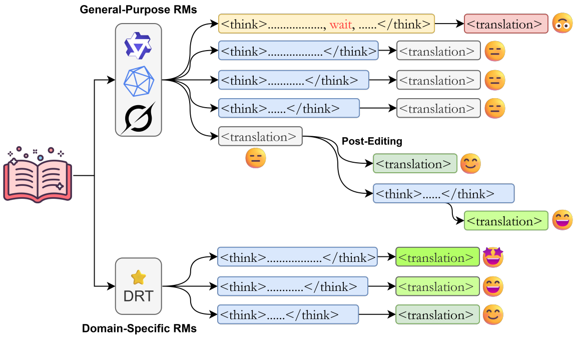


Figure 1: Illustration of the effectiveness of test-time scaling in reasoning models for machine translation. (1) TTS for general-purpose RMs yields only a small initial performance gain, but quickly plateauing as increased inference cost. (2) Forcing RMs to reason beyond their natural stopping point degrades quality by introducing noise. (3) In contrast, TTS becomes effective when applied to RMs specifically developed for MT. (4) TTS shows improvements in post-editing workflows. All these highlight TTS's value in MT lies in task-specialized models and multi-step self-correction, rather than as a robust strategy for enhancing single-pass translation with general-purpose RMs.

1 Introduction

Large language models (LLMs) have dramatically advanced machine translation (MT), evolving from statistical and neural paradigms to systems capable of handling diverse languages, domains, and complexities with unprecedented accuracy (Lyu et al., 2023; Kocmi et al., 2024; Zhu et al., 2024; Cui et al., 2025; Hendy et al., 2023). Recent developments in Reasoning Models (RMs), models designed to incorporate structured reasoning processes like Chain-of-Thought (CoT), have further transformed MT by reframing it as a cognitive task requiring contextual analysis, cultural adaptation, and self-reflection (Liu et al., 2025). For instance, RMs can resolve ambiguities in stylized texts and maintain coherence across documents, thereby outperforming traditional LLMs in semantically demanding scenarios (Ye et al., 2025).

Test-time scaling (TTS) has emerged as a transformative approach for enhancing model performance, which allocates additional computational resources during inference to enhance performance without requiring model retraining or parameter expansion (Snell et al., 2024). The effectiveness of TTS has been particularly pronounced for RMs such as DeepSeek-R1 (Guo et al., 2025), Gemini 2.5 (Comanici et al., 2025), and OpenAI's o-Series (Jaech et al., 2024; OpenAI, 2025), which have achieved

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2902–2917
March 24-29, 2026 ©2026 Association for Computational Linguistics

breakthrough performance on challenging benchmarks by extending their reasoning chains. Moreover, relatively small RMs have demonstrated impressive results on mathematical and coding tasks through TTS (Muennighoff et al., 2025; Li et al., 2025), suggesting that inference-time computation can partially compensate for limited model capacity.

Nevertheless, applying TTS to RMs for MT introduces distinct challenges and untapped potential that warrant deeper exploration. Unlike math or coding tasks, where correctness can often be objectively determined, MT demands not only linguistic accuracy but also reasoning over cultural nuances, domain-specific terminology, and long-range dependencies, areas where unstructured compute scaling may yield diminishing returns. Moreover, interventions like forced extrapolation (e.g., inserting "`wait`" tokens to extend reasoning) could disrupt natural deliberation, potentially introducing noise. In post-editing (PE) contexts, where models refine their own drafts, TTS might unlock iterative improvements, though this demands rigorous testing across varied benchmarks.

This paper investigates these open questions by distinguishing between two TTS workflows: *Direct Translation* (single-pass CoT scaling) and *Post-Editing* (compute-scaled self-correction). We structure our investigation through three research questions:

- **RQ1: How effective is test-time scaling for MT?** We examine whether increased inference computation reliably boosts translation quality across general-purpose and fine-tuned MT-specific RMs.
- **RQ2: Does extrapolation by inserting "`wait`" forcibly help?** We investigate if overriding models' natural stopping points, which further scales up the inference computation, enhances or hinders performance.
- **RQ3: Does test-time scaling work in post-editing?** We evaluate TTS in self-correction scenarios, assessing its role in refining initial translations when being allocated with specific compute budgets.

To address these questions, we assemble a comprehensive array of MT benchmarks encompassing literary, biomedical, cultural, commonsense, constrained terminology, and retrieval-augmented domains. We assess 12 RMs, spanning open-source series (Qwen-3 (Yang et al., 2025), Cogito (Deep Cogito, 2025), DRT (Wang et al., 2025a)) and the proprietary Grok-3-Mini.

Our key contributions and findings are as follows:

- We demonstrate that for general-purpose RM, TTS provides limited and inconsistent benefits for direct machine translation. After small initial improvements at very low budgets, performance plateaus across metrics and datasets, indicating that "more thinking" alone is not a robust path to better translations.
- We show that the effectiveness of TTS is unlocked by domain-specific fine-tuning, which aligns the model's reasoning process with task requirements. For DRT models fine-tuned on specific domain data, performance improves with budget on in-domain tasks and saturates once models naturally stop increasing their internal token usage, suggesting an emergent alignment between optimal reasoning depth and task demands. This alignment largely disappears out of the domain.
- We find that forcing a model to reason by inserting a single "`wait`" beyond its natural stopping point consistently degrades translation quality, highlighting the importance of the model's intrinsic deliberation process.
- We establish that TTS is highly effective in a post-editing context, in which the inference cost is higher than the cost of direct translation. TTS turns self-correction into a reliably beneficial process.

These findings provide implications for deploying RMs in production MT systems and highlight the critical interplay between model capacity, task-specific training, and inference-time computation in determining when and how test-time scaling benefits translation quality.

2 Experimental Setup

2.1 Datasets

To comprehensively evaluate the reasoning capabilities and scaling properties of models at test time, we curated a diverse suite of eight machine translation benchmarks. These datasets span multiple domains, granularities, and languages, targeting a wide spectrum of reasoning challenges (Table 1).

For tasks requiring deep contextual and stylistic understanding, we use three literary benchmarks: the document-level **WMT24-Literary** (Wang et al., 2024b), paragraph-level **LitEval-Corpus** (Zhang et al., 2025), and sentence-level **MetaphorTrans** (Wang et al., 2025a). These datasets are rich in complex linguistic phenomena, cultural references, and figurative language (similes and metaphors), demanding sophisticated reasoning to preserve literary style and meaning. Similarly, the **WMT23/24-Biomedical** (Neves et al., 2023, 2024) benchmark tests reasoning within a specialized domain, demanding accurate translation of technical terminology from PubMed abstracts.

To probe more targeted reasoning abilities, we incorporate four specialized datasets. **CAMT** (Yao et al., 2024) assesses cross-cultural reasoning on expressions requiring cultural adaptation. **Commonsense-MT** (He et al., 2020) comprises subsets targeting lexical, contextless syntactic, and contextual syntactic ambiguities, each requiring commonsense reasoning. **RTT** (Zhang et al., 2023) evaluates constrained reasoning by requiring models to correctly translate specific terminology under highly constrained conditions. Lastly, **RAG-Trans** (Wang et al., 2024a) examines a model's capacity to reason over and integrate retrieved external evidence into its translation. Collectively, these benchmarks provide a rigorous and multifaceted framework for analyzing the effects of scaling on translation reasoning.

2.2 Models

Our evaluation encompasses 12 RMs, including 11 open-source models from three distinct families and one proprietary model for comparison. The open-source models investigated are as follows:

- **Qwen-3**: Six models from this family were selected, with parameter sizes of 0.6B, 1.7B, 4B, 8B, 14B, and 32B (Yang et al., 2025). These models are hybrid reasoning LLMs that support seamless switching between a standard generation mode and a deliberative reasoning mode.
- **Cogito**: Two models, sized 3B and 8B, were included (Deep Cogito, 2025). Cogito-3B and Cogito-8B are trained on top of Llama-3.2-3B and Llama-3.1-8B (Grattafiori et al., 2024), respectively, and similarly implement hybrid reasoning capabilities with controllable switching between generation and reasoning modes.
- **DRT**: Three models from this family were evaluated. These models are fine-tuned from existing LLMs using the training set of MetaphorTrans (Wang et al., 2025a). Specifically, DRT-7B, DRT-8B, and DRT-14B are built upon Qwen2.5-7B-Instruct (Yang et al., 2024), Llama-3.1-8B-Instruct, and Qwen2.5-14B-Instruct respectively.

In addition to the open-source models, we included the proprietary model **Grok-3-Mini**. This model provides a tunable `reasoning_effort` parameter (equivalent to reasoning budget but can only be set to `low` or `high`) to control the amount of deliberation performed prior to generating a response.

2.3 Evaluation Metrics

We assess translation quality using a suite of automatic metrics, encompassing both reference-based and reference-free approaches, alongside a specialized LLM-based judge for literary texts.

COMET Metrics. For a standardized assessment, we employ two variants from the COMET framework (Rei et al., 2020): the reference-based **COMET-22** (Rei et al., 2022a) and the reference-free **COMETKiwi-22** (Rei et al., 2022b).

LLM as Judge. For LLM-based evaluation, we employ `Gemini-2.0-Flash`. We first define two general-purpose metrics, **Gemini Reference-Based (GRB)** and **Gemini Reference-Free (GRF)**, which provide a quality score on a 0-100 scale. Furthermore, for the specific challenges of literary translation, we follow Wang et al. (2025a) and apply the **Gemini Evaluation with Anchors (GEA)** metric exclusively to the three literary benchmarks. This specialized metric assesses nuances like style and expressiveness, and we collect scores at two levels of granularity: $\text{GEA}_{100} \in [0, 100]$ and $\text{GEA}_5 \in 1, \ldots, 5$. The evaluation prompts are adapted from Kocmi and Federmann (2023) and Wang et al. (2025a), are illustrated in Appendix A.

2.4 Budget Forcing

We regulate test-time reasoning with a logits processor that enforces a *thinking-token budget* inside a `<think>…</think>` span. While in this span, the processor counts tokens, softly encourages closure near 95% of the budget by upweighting newline and `</think>`, then deterministically emits a newline (penultimate step) and `</think>` (final step) at the budget limit before continuing normal answer decoding.

Conversely, to probe extrapolation, we optionally insert a single "`wait`" token if the model attempts to stop: specifically, after at least 5 thinking tokens, if `</think>` is the next token from the output of the argmax function and the budget is not yet exhausted,

Domain	Datasets	Granularity	Languages	Language Pair(s)	Sample Size
Literature	WMT24-Literary (Wang et al., 2024b)	Document-level	ZH, DE, RU	3	43
	MetaphorTrans (Wang et al., 2025a)	Sentence-level	ZH, EN	1	2000
	LitEval-Corpus (Zhang et al., 2025)	Paragraph-level	ZH, EN, DE	4	187
Biomedical	WMT24-Biomedical (Neves et al., 2024)	Document-level	EN, DE, ES, FR, IT, PT, RU	12	600
	WMT23-Biomedical (Neves et al., 2023)	Document-level			585
Culture	CAMT (Yao et al., 2024)	Sentence-level	EN, ES, FR, HI, TA, TE, ZH	7	6948
Commonsense	Commonsense-MT (He et al., 2020) (Lexical Ambiguity)	Sentence-level	ZH, EN	1	400
	Commonsense-MT (Contextless Syntactic Ambiguity)	Sentence-level			450
	Commonsense-MT (Contextual Syntactic Ambiguity)	Sentence-level			350
Terminology	RTT (Zhang et al., 2023)	Sentence-level	EN, DE	2	100
Misc.	RAGTrans (Wang et al., 2024a)	Sentence-level	ZH, EN	1	1999

Table 1: Overview of the MT benchmarks used in our evaluation.

we override the next token to "`wait`" once and resume unconstrained decoding. We insert "`wait`" at most once.

2.5 Post-Editing

Post-editing (or self-correction) involves two-stage translation, i.e., stage one of direct translation and stage two that corrects or post-edits the direct translation, which enables models to review and refine their own outputs (Feng et al., 2025; Wang et al., 2024c; Li et al., 2024). We explore two prompting strategies to guide this self-correction process, with full details provided in Appendix D. The first is a standard PE prompt, which we term "No QS" (No Quality Score). It provides the model with only the source text and its own draft translation to be refined. The second is an enhanced prompt, "QS" (with Quality Score), which additionally includes a numerical quality score of the draft, calculated as the average of the GRB and GRF scores from the initial translation. This provides the model with an explicit signal about the quality of the translation it needs to correct, potentially guiding a more targeted reasoning process.

3 Results and Analysis

3.1 Effectiveness of Test-Time Scaling

General-Purpose Models Show Limited Gains from Increased Budget. Our initial investigation focused on the efficacy of test-time scaling for general-purpose RMs, including the Qwen-3 and Cogito families, as well as the proprietary Grok-3-Mini model. These models were evaluated "out-of-the-box" without any fine-tuning on our benchmark datasets. Figure 2 plots the average GRB scores (right axis) alongside the actual thinking tokens generated (left axis) for the Qwen-3 and Cogito model series. After a small initial performance gain when moving from a zero budget to a minimal budget (e.g., 100 tokens), the models' performance curves almost completely plateau in most cases. A critical observation emerges from the thinking-token curves: despite being allocated budgets up to 2000 tokens, most general-purpose models fail to utilize this capacity. The actual reasoning length typically saturates around 600 tokens. This suggests that these models reach a "reasoning ceiling" where they autonomously terminate the deliberation process, indicating that *simply allocating more computational steps does not enable the models to produce more refined or accurate translations when they lack specific task-related knowledge.*

This conclusion is further corroborated and nuanced by our analysis of the Grok-3-Mini model. We analyzed its performance using both reference-based (GRB) and reference-free (GRF) metrics, visualized in Figure 3a and Figure 3b, respectively. Both metrics reveal a highly inconsistent, dataset-dependent impact. For instance, while higher effort improves scores on CommonsenseMT-Lexical across both GRB (+0.450) and GRF (+0.376), it significantly degrades performance on CommonsenseMT-Contextless in both cases (-0.780 for GRB, -0.367 for GRF). Crucially, the average effect across all datasets is negligible and even flips its sign depending on the metric: the mean GRB delta is a slightly negative -0.064, while the mean GRF delta is a slightly positive +0.033. Given that both scores are on a 100-point scale, these near-zero average changes underscore that the potential benefits and drawbacks of increased computational effort effectively cancel each other out, leading to no reliable overall improvement.

Therefore, our analyses of both open-source models with varying budgets and a proprietary model with different effort levels converge on a single conclusion: *for general-purpose LLMs without specific in-domain training, test-time scaling is not a robust*

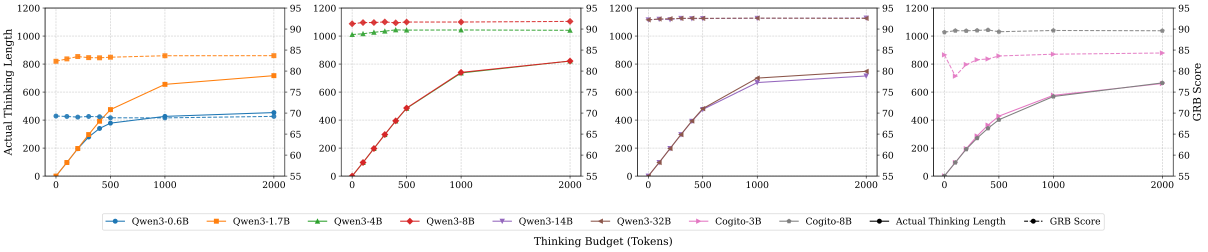
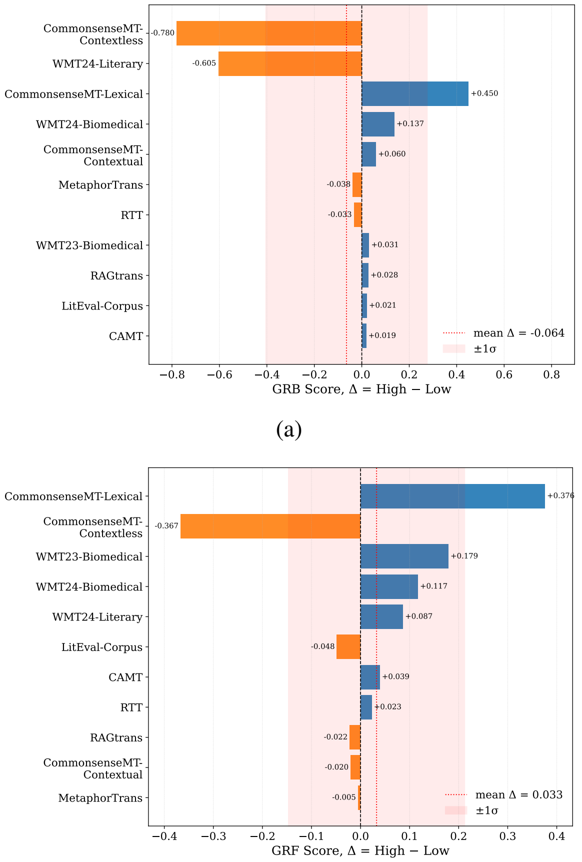


Figure 2: Average GRB scores and average actual thinking tokens of Qwen-3 and Cogito models across all datasets with varying thinking budgets.



(a)

(b)

Figure 3: Performance of Grok-3-mini across tasks, showing the difference between high- and low-effort reasoning. Subfigure (a) reports results under the GRB metric, and (b) shows results under GRF.

strategy for enhancing machine translation performance.

In-Domain Fine-Tuning Unlocks the Benefit of Test-Time Scaling. In contrast to the general-purpose models, the DRT models reveal that the effectiveness of test-time scaling is highly contingent on domain-specific training, which appears to create an efficient alignment between reasoning effort and performance. These models were fine-tuned on the training set of MetaphorTrans (in-domain), while LitEval-Corpus and WMT24-Literary serve as related but out-of-domain literary benchmarks. Figure 4 visualizes the translation quality (GEA score, right y-axis) and the actual number of thinking tokens generated by the models (left y-axis).

On the in-domain MetaphorTrans task, we observe a clear and consistent positive correlation between the thinking budget and translation performance. We consider the model's natural stopping point, where the 'Actual tokens' curve plateaus around 500 tokens, as the realistic baseline. As the thinking budget increases from 100 to this limit, both the number of generated tokens and the GEA scores steadily rise. This monotonic improvement indicates that the model is performing valid, necessary reasoning steps.

However, beyond a 500-token budget, a critical pattern emerges: the models stop generating more thinking tokens, and concurrently, their performance plateaus. We hypothesize that the fine-tuning successfully aligns the model's reasoning behavior with the effective reasoning boundary of the task.

This efficient alignment vanishes on the other out-of-domain literary translation tasks. The most striking counterexample is the document-level WMT24-Literary task. Here, the actual thinking tokens continue to scale almost linearly with the budget, indicating the models are using the provided extra capacity to "think" longer.[1] Yet, this extended reasoning does not translate into better performance; the GEA scores remain erratic and show no consistent improvement. This disconnect suggests the models are engaged in unproductive or unfocused reasoning, "spinning their wheels" without the spe-

[1] We attribute this distinct behavior to text granularity: unlike the out-of-domain paragraph-level tasks (LitEval-Corpus) where limited context leads models to exhaust reasoning paths and saturate early, the extensive context in document-level translation allows models to continuously expand their reasoning loops to consume the available budget.

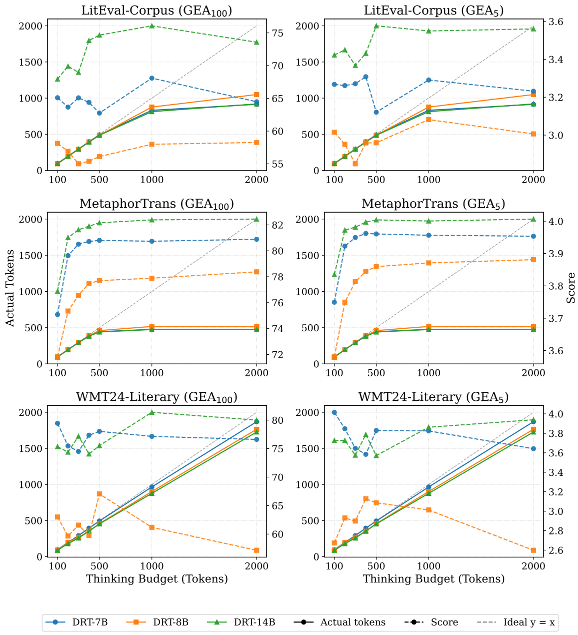


Figure 4: Performance (dashed lines, right axis) and actual generated thinking tokens (solid lines, left axis) of DRT models across 3 literary translation tasks.

cialized knowledge required for this different type of literary translation. This dichotomy underscores our central argument: *test-time scaling is most effective when fine-tuning has equipped a model with not only domain-specific knowledge but also an efficient strategy for how and when to apply it.*

3.2 Forced Extrapolation of Reasoning Degrades Performance

Building on the observation that models possess a natural reasoning length, we next address RQ2: Does performance improve if we force a model to think longer? We test this by applying the "`wait`" token extrapolation method, detailed in Section 2.4, to prompt continued reasoning when the models are about to stop the thinking process.

The results, averaged across all datasets for the Qwen-3 and Cogito models, are presented in Table 2. The findings are unambiguous. First, as evidenced by the "Thinking Length" columns, the intervention was effective in its primary goal: it consistently and significantly extended the models' reasoning chains, often by over 100-200 tokens.

However, this artificially prolonged reasoning process did not translate into better translations. In fact, it was overwhelmingly detrimental. Across all four metrics (COMET, COMETKiwi, GRB, and GRF), the "`wait`" token intervention consistently reduces performance: under both the 1000- and 2000-token budgets, 55 of the 64 metric scores across eight models dropped after forced extrapolation. While there are a few isolated instances of negligible score increases in one metric (e.g., Qwen3-4B on GRB with a 2000-token budget), these are exceptions that are contradicted by decreases in other metrics for the same model.

This leads to a clear conclusion: a model's decision to terminate its reasoning chain is a meaningful signal. It indicates that the model has reached what it considers to be a sufficient state of deliberation for the given task. *Forcing the model to continue to reason beyond its own stopping point appears to introduce noise, repetition, or less relevant reasoning steps, which ultimately harms the quality of the final translation.* In short, we find that forced extrapolation is a counterproductive strategy for improving translation quality.

3.3 Test-Time Scaling is Effective for Post-Editing

Finally, we investigate RQ3 by evaluating the effectiveness of test-time scaling in a self-correction post-editing scenario. For this task, we define the baseline as the translation generated by each Qwen-3 model with a zero thinking budget in our prior experiments. Subsequently, we task the same model with refining its own translation, applying thinking budgets of 0, 500, and 1000 tokens. The detailed results are presented in Table 12 and Table 13, with trends visualized in Figure 5.

In a striking contrast to its ineffectiveness in direct translation, *test-time scaling proves to be a highly effective strategy for post-editing, reliably elevating translation quality above the original baseline for most models.* The effect is most pronounced for mid-sized models, as shown in both the GRB (Figure 5a) and GRF (Figure 5b) plots. For models in the 1.7B to 14B parameter range, applying post-editing with a zero budget often yields results that are similar to or worse than the original translation. Increasing the budget to 500 or 1000 tokens consistently pushes performance significantly above this baseline, demonstrating that a thinking budget is crucial for turning self-correction into a reliably beneficial process.

However, this scaling trend does not hold for the extremes of the model family. The smallest model, Qwen3-0.6B, displays erratic behavior, with its performance fluctuating without clear improvement as the budget increases. Conversely, the largest model, Qwen3-32B, already surpasses the baseline with

Model	Budget	Thinking Length		COMET		COMETKiwi		GRB		GRF	
		Before	After	Before	After	Before	After	Before	After	Before	After
Qwen3-0.6B	1000	426	519	**0.6959**	0.6904	**0.6155**	0.6026	**68.8577**	68.6333	**67.2944**	67.1203
	2000	454	556	**0.6895**	0.6894	**0.6087**	0.6016	**69.2139**	69.0237	**67.6465**	67.3326
Qwen3-1.7B	1000	655	820	**0.7687**	0.7475	**0.6851**	0.6543	**83.6481**	83.5185	**83.1170**	82.7297
	2000	717	987	**0.7645**	0.7496	**0.6762**	0.6586	**83.6647**	83.4715	**83.1076**	82.7472
Qwen3-4B	1000	737	873	**0.7914**	0.7738	**0.7092**	0.6835	**89.7781**	89.5598	**89.7522**	89.5203
	2000	822	1098	**0.7871**	0.7784	**0.7005**	0.6862	89.7023	**90.0098**	89.6536	**89.7275**
Qwen3-8B	1000	741	878	**0.7979**	0.7865	**0.7180**	0.6954	91.6706	**91.7601**	**91.8085**	91.7027
	2000	821	1117	**0.7965**	0.7860	**0.7118**	0.6951	**91.8186**	91.6111	91.8123	**91.9008**
Qwen3-14B	1000	668	812	**0.7992**	0.7924	**0.7184**	0.7022	92.5609	**92.5677**	**92.5242**	92.5148
	2000	715	955	**0.7966**	0.7908	**0.7125**	0.7020	**92.6259**	92.5421	**92.5617**	92.5392
Qwen3-32B	1000	700	835	**0.8025**	0.7828	**0.7233**	0.6957	**92.6026**	92.5537	**92.6328**	92.5322
	2000	748	992	**0.7993**	0.7786	**0.7174**	0.6937	92.5393	**92.6689**	92.6698	**92.7696**
Cogito-3B	1000	577	695	**0.7071**	0.7063	**0.6313**	0.6310	**84.0098**	83.8203	**83.1502**	82.5440
	2000	661	844	0.7040	**0.7049**	0.6324	**0.6331**	**84.2981**	83.9921	**82.4356**	81.8855
Cogito-8B	1000	568	697	**0.7678**	0.7658	**0.6824**	0.6807	**89.6573**	89.3849	**89.2655**	88.8525
	2000	665	882	**0.7687**	**0.7687**	**0.6823**	0.6809	**89.5962**	89.4339	**89.3591**	89.2452

Table 2: Effect of forcibly inserting a "`wait`" token to extend the reasoning process. The "Before" columns show standard generation, while "After" shows results from the intervention.

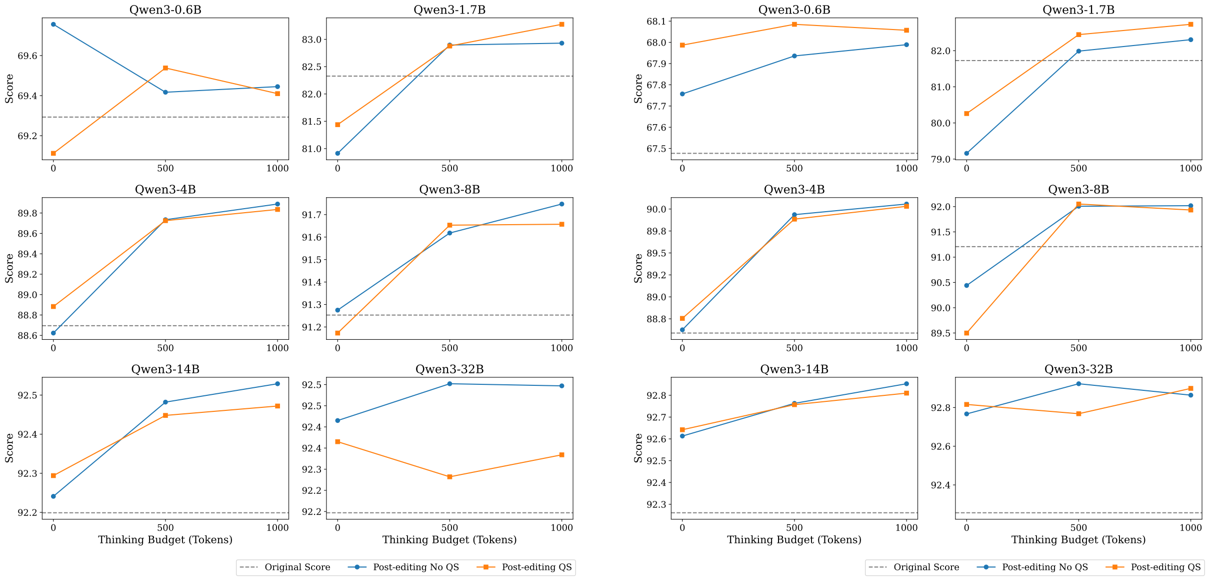
a zero-budget correction, and additional thinking time provides no further gains, suggesting it performs near its peak without extended deliberation.

A comparison of the two prompting strategies further highlights the importance of the thinking budget. At a zero-token budget, the "QS" prompt (with quality score) generally underperforms the "No QS" prompt. However, once the budget is increased to 500 or 1000 tokens, their performances converge and become nearly indistinguishable. This demonstrates that while prompting strategy matters, it is the allocation of a computational budget that is the key factor for post-editing to reliably improve upon the initial translation.

4 Related Work

Machine Translation with Large Language Models The application of LLMs to machine translation has witnessed a paradigm shift. Foundational research demonstrated that general-purpose models, such as GPT-3, possess remarkable few-shot translation capabilities, challenging traditional supervised systems (Brown et al., 2020). Subsequent empirical studies systematically evaluated these capabilities, revealing that while LLMs excel in high-resource languages and specific domains, they often require careful adaptation to match state-of-the-art baselines (Hendy et al., 2023; Zhu et al., 2024). To address these limitations, researchers have developed sophisticated in-context learning strategies (Agrawal et al., 2023) and diverse prompting techniques (Vilar et al., 2023). Notably, recent work by Wu et al. (2025) highlights the efficacy of simple re-translation prompts over complex reasoning in general LLMs, suggesting that iterative refinement can significantly boost performance without elaborate chains of thought. Concurrently, the field is moving towards open-weight models that balance broad multilingual competence with translation specialization. Seed-X (Cheng et al., 2025) introduces a 7B-parameter open-source family of translation-oriented LLMs trained on large-scale monolingual and bilingual corpora across 28 languages, achieving performance competitive with closed-source systems such as GPT-4o (Hurst et al., 2024) and Gemini-2.5 (Comanici et al., 2025). Similarly, Hunyuan-MT (Zheng et al., 2025) develops 7B-parameter models: Hunyuan-MT and Hunyuan-MT-Chimera, the latter integrates multiple outputs under a "slow thinking" paradigm to yield higher-quality translations, and ranks first in the WMT2025 shared task across 30 of 31 directions (Kocmi et al., 2025). Tower+ (Rei et al., 2025) addresses the trade-off between translation specialization and general-purpose ability by combining continued pretraining, supervised fine-tuning, preference optimization, and reinforcement learning with verifiable rewards.

Machine Translation with Reasoning Models Recent research has explored how RMs can be adapted to MT, particularly in linguistically and culturally challenging domains. Wang et al. (2025a) propose Deep Reasoning Translation (DRT), which



(a) GRB scores for post-editing across Qwen-3 models.

(b) GRF scores for post-editing across Qwen-3 models.

Figure 5: Effectiveness of test-time scaling in post-editing scenario.

leverages long chain-of-thought (CoT) reasoning within a multi-agent framework to tackle similes and metaphors in English–Chinese literary translation. The resulting models surpass standard LLMs by synthesizing long-thought training data. Building on this, DeepTrans (Wang et al., 2025b) applies reinforcement learning (RL) with carefully designed reward functions targeting both translation fidelity and reasoning quality, showing that RL without labeled pairs can significantly boost performance. ExTrans (Wang et al., 2025c) complements this direction with an exemplar-enhanced RL approach that employs a stronger RM (DeepSeek-R1) as a reward reference. ExTrans achieves state-of-the-art results in English–Chinese literary MT, and its multilingual extension (mExTrans) scales effectively to 90 directions with lightweight reward modeling. Beyond literary MT, R1-T1 (He et al., 2025) generalizes reasoning-based MT by modeling six CoT templates inspired by human translator strategies. Through RL, it enables self-evolving reasoning trajectories, improving performance across diverse domains and low-resource languages.

Test-Time Scaling Recent studies have examined the potential of test-time scaling. Tan et al. (2025) propose a best-of-N reranking framework where multiple translation candidates are generated and the best one is selected using a quality estimation model. They show that smaller models can, through TTS, match or even surpass larger model. For example, a 14B model with $N \approx 8$ achieves parity with a 72B model at $N = 1$ while requiring substantially less GPU memory. Beyond MT, Son et al. (2025) analyze TTS in multilingual mathematical reasoning, finding that outcome and process reward modeling, as well as budget forcing, yield notable gains in English but limited improvements across 55 languages, highlighting cross-lingual fragility. Tran et al. (2025) study low-resource reasoning tasks and propose English-pivoted CoT generation, where reasoning occurs in English before producing final answers in the target language, yielding substantial accuracy improvements. Yong et al. (2025) further study cross-lingual reasoning with English-centric RMs, finding that scaling inference budgets with long CoTs improves multilingual mathematical reasoning and even allows smaller models to outperform larger baselines, but also highlighting language-mixing behaviors and weaker generalization to cultural commonsense domains.

5 Conclusion

In this work, we systematically explored the application of test-time scaling (TTS) to reasoning models (RMs) for machine translation (MT), addressing three core research questions through extensive experiments across diverse benchmarks, models, and evaluation metrics.

Our findings reveal that TTS offers limited value for general-purpose RMs in direct translation tasks, where performance quickly plateaus after minimal initial gains, underscoring that additional inference-time computation alone cannot compensate for a lack of task-aligned reasoning strategies. In contrast, domain-specific fine-tuning emerges as a pivotal enabler, allowing TTS to yield consistent improvements on in-domain tasks until models reach their natural reasoning depth, beyond which further scaling provides no benefit. This highlights an emergent efficiency in fine-tuned models, where optimal deliberation aligns with task demands, though such alignment erodes out-of-domain. Furthermore, forcibly extending reasoning via interventions like "`wait`" tokens consistently degrades quality, emphasizing the importance of respecting a model's intrinsic stopping points. Finally, TTS proves particularly potent in post-editing scenarios, transforming self-correction into a reliable mechanism for refining initial drafts, especially for mid-sized models when paired with adequate budgets.

The implications of this work are twofold. First, for practitioners, simply allocating more inference compute to general-purpose models is an inefficient path to better translations. Instead, resources are better invested in targeted fine-tuning, which aligns the model's reasoning capabilities with specific task demands. Second, our results suggest that the most promising application of TTS in MT is not in direct, single-pass translation but in multi-stage workflows, such as a rapid initial draft followed by a more deliberate, computationally-intensive self-correction phase. Future work could explore more dynamic budget allocation strategies and extend to hybrid TTS approaches integrated with external tools like retrieval-augmented generation.

Limitations

While our study provides a comprehensive analysis of test-time scaling in machine translation, we acknowledge several limitations that frame the scope of our conclusions and suggest avenues for future research.

First, our investigation, while encompassing 12 different models, is primarily focused on open-source RM families and a single, smaller proprietary model. The performance characteristics and scaling behaviors of the largest, state-of-the-art proprietary models (e.g., Gemini-2.5-Pro) may differ from our observations. Furthermore, the linguistic diversity of our benchmarks is largely centered around English or Chinese as either a source or target language. Consequently, our findings on the effectiveness of TTS, particularly the interplay with fine-tuning, may not generalize directly to low-resource languages where the reasoning challenges could be substantially different.

Second, our evaluation methodology relies exclusively on automatic and LLM-based metrics. Although we employed a suite of reference-based, reference-free, and specialized LLM-judge metrics to ensure robustness, this approach lacks the nuance of human evaluation. A human assessment would be invaluable for validating our findings, especially on the literary and cultural benchmarks where subtle aspects of style, tone, and appropriateness are critical and may not be fully captured by our current metrics. The potential biases inherent in LLM-as-a-judge frameworks also represent a confounding factor.

Finally, our study implements test-time scaling through a specific budget-forcing mechanism and a simple "`wait`" token intervention for extrapolation. Other methods for encouraging or extending deliberation, such as alternative prompting strategies or more complex reasoning frameworks, were not explored and could yield different outcomes. Additionally, our analysis is primarily quantitative; we did not perform a qualitative analysis of the content within the models' reasoning chains. A deeper investigation into what the models are "thinking" could provide valuable insights into why performance plateaus for general-purpose models or why forced extrapolation leads to degradation.

Acknowledgments

The work has received funding from the Digital Europe Programme under grant agreement No 101195233 (OpenEuroLLM). The authors wish to acknowledge CSC - IT Center for Science, Finland, and LUMI supercomputers, owned by the EuroHPC Joint Undertaking, for providing computational resources. Shaoxiong Ji received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

References

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

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.

Shanbo Cheng, Yu Bao, Qian Cao, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu, Jingwen Chen, Zhichao Huang, Tao Li, Yifu Li, Huiying Lin, Sitong Liu, Ningxin Peng, Shuaijie She, Lu Xu, Nuo Xu, Sen Yang, and 7 others. 2025. Seed-x: Building strong multilingual translation llm with 7b parameters. *Preprint*, arXiv:2507.13618.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5420–5443, Albuquerque, New Mexico. Association for Computational Linguistics.

Deep Cogito. 2025. Cogito v1 Preview Introducing IDA as a path to general superintelligence. `https://www.deepcogito.com/research/cogito-v1-preview`. Accessed: 2025-09-09.

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2025. TEaR: Improving LLM-based machine translation with systematic self-refinement. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3922–3938, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

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

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

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

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mkadry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alexander Kirillov, Alex Nichol, Alex Paino, and 397 others. 2024. Gpt-4o system card. *ArXiv*, abs/2410.21276.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, and 242 others. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

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.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Dacheng Li, Shiyi Cao, Chengkun Cao, Xiuyu Li, Shangyin Tan, Kurt Keutzer, Jiarong Xing, Joseph E Gonzalez, and Ion Stoica. 2025. S*: Test time scaling for code generation. *arXiv preprint arXiv:2502.14382*.

Xinnuo Li, Yunxiang Zhang, and Lu Wang. 2024. Improving language model self-correction capability with meta-feedback. *OpenReview*.

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

Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F Wong, Siyou Liu, and Longyue Wang. 2023. A paradigm shift: The future of machine translation lies with large language models. *arXiv preprint arXiv:2305.01181*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Mariana Neves, Cristian Grozea, Philippe Thomas, Roland Roller, Rachel Bawden, Aurélie Névéol, Steffen Castle, Vanessa Bonato, Giorgio Maria Di Nunzio, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, and Antonio Jimeno Yepes. 2024. Findings of the WMT 2024 biomedical translation shared task: Test sets on abstract level. In *Proceedings of the Ninth Conference on Machine Translation*, pages 124–138, Miami, Florida, USA. Association for Computational Linguistics.

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, and Cristian Grozea. 2023. Findings of the WMT 2023 biomedical translation shared task: Evaluation of ChatGPT 3.5 as a comparison system. In *Proceedings of the Eighth Conference on Machine Translation*, pages 43–54, Singapore. Association for Computational Linguistics.

OpenAI. 2025. Openai o3 and o4-mini system card. System card, OpenAI. Published April 16, 2025.

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

Ricardo Rei, Nuno M Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André FT Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. *arXiv preprint arXiv:2506.17080*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

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

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Guijin Son, Jiwoo Hong, Hyunwoo Ko, and James Thorne. 2025. Linguistic generalizability of test-time scaling in mathematical reasoning. *arXiv preprint arXiv:2502.17407*.

Shaomu Tan, Ryosuke Mitani, Ritvik Choudhary, and Toshiyuki Sekiya. 2025. Investigating test-time scaling with reranking for machine translation. *arXiv preprint arXiv:2509.19020*.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang D Nguyen. 2025. Scaling test-time compute for low-resource languages: Multilingual reasoning in llms. *arXiv preprint arXiv:2504.02890*.

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.

Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2025a. DRT: Deep reasoning translation via long chain-of-thought. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6770–6782, Vienna, Austria. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Yingxue Zhang, and Jie Zhou. 2024a. Retrieval-augmented machine translation with unstructured knowledge. *arXiv preprint arXiv:2412.04342*.

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

Jiaan Wang, Fandong Meng, and Jie Zhou. 2025c. Extrans: Multilingual deep reasoning translation via exemplar-enhanced reinforcement learning. *arXiv preprint arXiv:2505.12996*.

Longyue Wang, Siyou Liu, Chenyang Lyu, Wenxiang Jiao, Xing Wang, Jiahao Xu, Zhaopeng Tu, Yan Gu, Weiyu Chen, Minghao Wu, Liting Zhou, Philipp Koehn, Andy Way, and Yulin Yuan. 2024b. Findings of the WMT 2024 shared task on discourse-level literary translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 699–700, Miami, Florida, USA. Association for Computational Linguistics.

Yutong Wang, Jiali Zeng, Xuebo Liu, Fandong Meng, Jie Zhou, and Min Zhang. 2024c. TasTe: Teaching large language models to translate through self-reflection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6144–6158, Bangkok, Thailand. Association for Computational Linguistics.

Di Wu, Seth Aycock, and Christof Monz. 2025. Please translate again: Two simple experiments on whether human-like reasoning helps translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20424–20440, Suzhou, China. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

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

Yongshi Ye, Biao Fu, Chongxuan Huang, Yidong Chen, and Xiaodong Shi. 2025. How well do large reasoning models translate? a comprehensive evaluation for multi-domain machine translation. *arXiv preprint arXiv:2505.19987*.

Zheng-Xin Yong, M Farid Adilazuarda, Jonibek Mansurov, Ruochen Zhang, Niklas Muennighoff, Carsten Eickhoff, Genta Indra Winata, Julia Kreutzer, Stephen H Bach, and Alham Fikri Aji. 2025. Crosslingual reasoning through test-time scaling. *arXiv preprint arXiv:2505.05408*.

Huaao Zhang, Qiang Wang, Bo Qin, Zelin Shi, Haibo Wang, and Ming Chen. 2023. Understanding and improving the robustness of terminology constraints in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6029–6042, Toronto, Canada. 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. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10961–10988, Albuquerque, New Mexico. Association for Computational Linguistics.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. Hunyuan-mt technical report. *arXiv preprint arXiv:2509.05209*.

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 Evaluation Prompts

GRB Prompt:

> Score the following translation from {src_lang} to {tgt_lang} with respect to the human reference on a continuous scale from 0 to 100, where a score of zero means "no meaning preserved" and score of one hundred means "perfect preservation of meaning, with faithfulness, expressiveness, and elegance".
> Only output the score number.
>
> {src_lang} source: "{src_text}"
> {tgt_lang} human reference: "{tgt_text}"
> {tgt_lang} translation: "{hyp_text}"
>
> Score:

GRF Prompt:

```
Score the following translation from {src_lang} to {tgt_lang} on a continuous scale from 0 to 100, where a score of zero means "no meaning preserved" and score of one hundred means "perfect preservation of meaning, with faithfulness, expressiveness, and elegance".
Only output the score number.

{src_lang} source: "{src_text}"
{tgt_lang} translation: "{hyp_text}"

Score:
```

GEA$_{100}$ Prompt:

```
Please evaluate the following {tgt_lang} translation of an {src_lang} text. Rate the translation on a scale of 0 to 100, where:

- 10 points: Poor translation; the text is somewhat understandable but contains significant errors and awkward phrasing that greatly hinder comprehension for a {tgt_lang} reader.
- 30 points: Fair translation; the text conveys the basic meaning but lacks fluency and contains several awkward phrases or inaccuracies, making it challenging for a {tgt_lang} reader to fully grasp the intended message.
- 50 points: Good translation; the text is mostly fluent and conveys the original meaning well, but may have minor awkwardness or slight inaccuracies that could confuse a {tgt_lang} reader.
- 70 points: Very good translation; the text is smooth and natural, effectively conveying the intended meaning, but may still have minor issues that could slightly affect understanding for a {tgt_lang} reader.
- 90 points: Excellent translation; the text is fluent and natural, conveying the original meaning clearly and effectively, with no significant issues that would hinder understanding for a {tgt_lang} reader.

Please only output the score number.
```

GEA$_5$ Prompt:

```
Please evaluate the following {tgt_lang} translation of an {src_lang} text. Rate the translation on a scale of 0 to 5, where:

- 1 point: Poor translation; the text is somewhat understandable but contains significant errors and awkward phrasing that greatly hinder comprehension for a {tgt_lang} reader.
- 2 points: Fair translation; the text conveys the basic meaning but lacks fluency and contains several awkward phrases or inaccuracies, making it challenging for a {tgt_lang} reader to fully grasp the intended message.
- 3 points: Good translation; the text is mostly fluent and conveys the original meaning well, but may have minor awkwardness or slight inaccuracies that could confuse a {tgt_lang} reader.
- 4 points: Very good translation; the text is smooth and natural, effectively conveying the intended meaning, but may still have minor issues that could slightly affect understanding for a {tgt_lang} reader.
- 5 points: Excellent translation; the text is fluent and natural, conveying the original meaning clearly and effectively, with no significant issues that would hinder understanding for a {tgt_lang} reader.

Please only output the score number.
```

B Evaluation Results of General-purpose Models

Tables 3–6 summarize the average performance of the General-purpose models across all datasets with respect to the COMET, COMETKiwi, GRB, and GRF metrics, respectively.

Model	Budget								
	0	100	200	300	500	1000	2000	Low	High
Cogito-3B	0.712	0.671	0.695	0.698	0.704	0.707	0.704		
Cogito-8B	0.765	0.762	0.764	0.763	0.762	0.768	0.769		
Qwen3-0.6B	0.702	0.694	0.699	0.695	0.695	0.696	0.689		
Qwen3-1.7B	0.760	0.757	0.764	0.763	0.764	0.769	0.764		
Qwen3-4B	0.789	0.785	0.789	0.790	0.791	0.791	0.787		
Qwen3-8B	0.801	0.800	0.799	0.800	0.799	0.798	0.797		
Qwen3-14B	0.805	0.807	0.803	0.806	0.805	0.799	0.797		
Qwen3-32B	0.804	0.804	0.802	0.803	0.802	0.802	0.799		
Grok-3-Mini								0.794	0.795

Table 3: Average COMET scores of general-purpose models across all datasets with varying thinking budgets.

Model	Budget								
	0	100	200	300	500	1000	2000	Low	High
Cogito-3B	0.634	0.599	0.619	0.625	0.629	0.631	0.632		
Cogito-8B	0.677	0.676	0.681	0.679	0.680	0.682	0.682		
Qwen3-0.6B	0.618	0.617	0.616	0.617	0.614	0.615	0.609		
Qwen3-1.7B	0.675	0.671	0.678	0.678	0.679	0.685	0.676		
Qwen3-4B	0.706	0.702	0.705	0.707	0.709	0.709	0.700		
Qwen3-8B	0.719	0.718	0.717	0.718	0.717	0.718	0.712		
Qwen3-14B	0.725	0.724	0.723	0.725	0.725	0.718	0.713		
Qwen3-32B	0.724	0.720	0.721	0.722	0.724	0.723	0.717		
Grok-3-Mini								0.701	0.701

Table 4: Average COMETKiwi scores of general-purpose models across all datasets with varying thinking budgets.

Model	Budget								
	0	100	200	300	500	1000	2000	low	high
Cogito-3B	82.546	78.016	79.590	80.571	81.176	81.497	81.937		
Cogito-8B	88.177	88.552	88.312	88.402	87.948	88.432	88.351		
Qwen3-0.6B	58.165	57.507	57.518	57.813	57.317	57.372	57.734		
Qwen3-1.7B	74.422	74.634	75.522	74.883	75.076	75.678	75.753		
Qwen3-4B	84.909	84.921	85.361	85.809	86.165	86.379	86.148		
Qwen3-8B	89.204	89.394	89.441	89.623	89.651	89.541	90.001		
Qwen3-14B	90.899	90.875	90.707	91.075	91.163	91.253	91.294		
Qwen3-32B	90.949	91.103	91.204	91.333	91.340	91.556	91.455		
Grok-3-Mini								92.529	92.451

Table 5: Average GRB scores of general-purpose models across all datasets with varying thinking budgets.

C Evaluation Results and Thinking Token Statistics of DRT Models on Literary Translation Tasks

Tables 7- 10 present the performance of the DRT models across three literary translation benchmarks with respect to the GEA_{100}, GEA_5, GRB, and GRF metrics, respectively. Table 11 shows the token statistics under different budgets. Figure 6 visualizes DRT models' translation quality (GRB&GRF score, right y-axis) and the actual number of thinking tokens (left y-axis). On the in-domain MetaphorTrans task, there is a positive growth of the quality scores as the thinking budget increases to around 300 tokens, then the score generally stabilizes. While on the other two out-of-domain tasks, the performance fluctuates and shows an overall downward trend as thinking tokens increase.

D Post-editing Prompts and Detailed Results

D.1 Prompts

We experiment with two variants of post-editing prompts: with and without an additional quality score (QS). Examples are provided below.

Post-editing with QS:

Model	Budget								
	0	100	200	300	500	1000	2000	low	high
Cogito-3B	80.928	77.468	80.178	81.680	81.577	81.843	80.632		
Cogito-8B	89.470	89.550	88.707	88.942	88.082	89.067	89.284		
Qwen3-0.6B	56.858	56.943	57.549	57.137	57.232	56.890	56.990		
Qwen3-1.7B	74.954	75.377	76.210	75.829	75.723	76.770	76.591		
Qwen3-4B	86.238	86.229	86.722	86.965	87.603	87.806	87.577		
Qwen3-8B	90.321	90.625	90.648	90.816	91.024	91.123	91.258		
Qwen3-14B	92.084	92.016	91.932	92.359	92.503	92.492	92.547		
Qwen3-32B	92.122	92.268	92.404	92.608	92.702	92.835	92.782		
Grok-3-Mini								93.494	93.594

Table 6: Average GRF scores of general-purpose models across all datasets with varying thinking budgets.

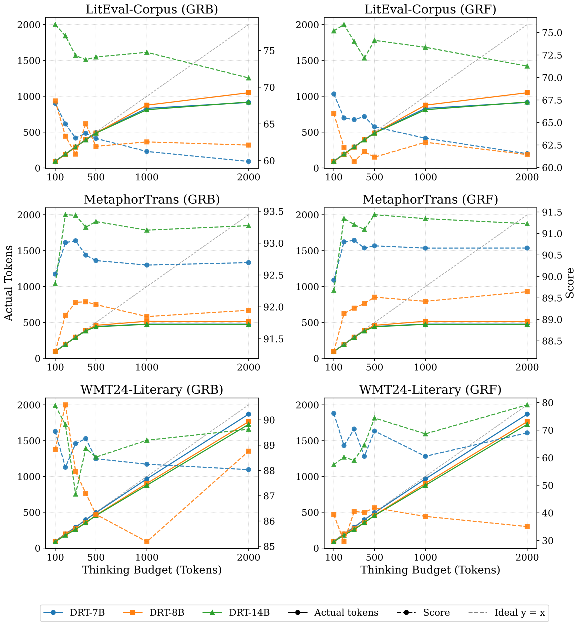


Figure 6: Performance and actual generated thinking tokens of DRT models across 3 literary translation tasks.

You are a professional translator, and your task is to refine the {tgt_lang} draft translation below based on the {src_lang} source text and its quality evaluation.
Please only provide me with the refined translation, without any additional explanations.
Source Text: {src_text}
Draft Translation: {hyp_text}
Quality Score: {quality_score}/100

Post-editing without QS:

You are a professional translator, and your task is to refine the {tgt_lang} draft translation below based on the {src_lang} source text.
Please only provide me with the refined translation, without any additional explanations.
Source Text: {src_text}
Draft Translation: {hyp_text}

Task	Model	Budget						
		100	200	300	400	500	1000	2000
MetaphorTrans	DRT-7B	75.08	79.63	80.52	80.72	80.81	80.74	80.89
	DRT-8B	71.81	75.36	76.57	77.48	77.70	77.89	78.38
	DRT-14B	76.88	81.02	81.62	81.90	82.16	82.39	82.45
LitEval-Corpus	DRT-7B	65.09	63.66	65.07	64.38	62.75	68.08	64.47
	DRT-8B	58.11	56.92	55.00	55.39	56.10	57.98	58.26
	DRT-14B	67.94	69.90	68.97	73.83	74.65	76.07	73.57
WMT24-Literary	DRT-7B	79.45	75.48	74.50	77.35	78.03	77.15	76.62
	DRT-8B	63.00	59.66	61.57	59.75	67.05	61.18	57.16
	DRT-14B	75.34	74.38	77.18	74.06	75.54	81.37	80.01

Table 7: GEA_{100} scores of DRT models on literary translation tasks with varying thinking budgets.

Task	Model	Budget						
		100	200	300	400	500	1000	2000
MetaphorTrans	DRT-7B	3.75	3.92	3.95	3.96	3.96	3.96	3.95
	DRT-8B	3.58	3.75	3.81	3.85	3.86	3.87	3.88
	DRT-14B	3.83	3.97	3.98	4.00	4.00	4.00	4.01
LitEval-Corpus	DRT-7B	3.27	3.26	3.27	3.31	3.12	3.29	3.23
	DRT-8B	3.02	2.95	2.85	2.96	2.96	3.08	3.01
	DRT-14B	3.42	3.45	3.37	3.43	3.58	3.55	3.56
WMT24-Literary	DRT-7B	4.02	3.85	3.65	3.58	3.83	3.83	3.64
	DRT-8B	2.67	2.93	2.90	3.13	3.09	3.01	2.60
	DRT-14B	3.73	3.73	3.58	3.79	3.57	3.86	3.94

Table 8: GEA_5 scores of DRT models on literary translation tasks with varying thinking budgets.

D.2 Detailed Results

Tables 12 and 13 report the GRB and GRF scores respectively.

Task	Model	Budget						
		100	200	300	400	500	1000	2000
MetaphorTrans	DRT-7B	92.51	93.01	93.04	92.81	92.73	92.66	92.69
	DRT-8B	91.30	91.87	92.07	92.08	92.04	91.85	91.95
	DRT-14B	92.36	93.45	93.44	93.25	93.34	93.20	93.27
LitEval-Corpus	DRT-7B	67.81	65.00	63.08	63.73	63.02	61.25	59.89
	DRT-8B	68.13	63.33	60.89	65.02	61.95	62.55	62.12
	DRT-14B	78.52	76.99	74.30	73.73	74.11	74.75	71.26
WMT24-Literary	DRT-7B	89.54	88.13	89.06	89.26	88.46	88.25	88.03
	DRT-8B	88.83	90.59	87.96	87.10	86.26	85.18	88.76
	DRT-14B	90.56	89.81	87.06	88.87	88.53	89.19	89.63

Table 9: GRB scores of DRT models on literary translation tasks with varying thinking budgets.

Task	Model	Budget						
		100	200	300	400	500	1000	2000
MetaphorTrans	DRT-7B	89.91	90.81	90.84	90.66	90.71	90.66	90.66
	DRT-8B	88.25	89.14	89.26	89.37	89.52	89.42	89.65
	DRT-14B	89.67	91.35	91.21	91.09	91.44	91.34	91.23
LitEval-Corpus	DRT-7B	68.17	65.50	65.32	65.66	64.52	63.25	61.53
	DRT-8B	66.01	62.22	60.67	61.75	61.15	62.80	61.43
	DRT-14B	75.18	75.88	73.99	72.16	74.13	73.36	71.27
WMT24-Literary	DRT-7B	76.06	64.40	70.39	60.50	69.66	60.51	69.00
	DRT-8B	39.38	29.54	40.48	40.19	41.83	38.72	35.03
	DRT-14B	57.46	60.20	59.01	64.58	74.38	68.64	79.16

Table 10: GRF scores of DRT models on literary translation tasks with varying thinking budgets.

Task	Model	Budget						
		100	200	300	400	500	1000	2000
MetaphorTrans	DRT-7B	96.50	195.74	295.37	387.48	443.83	476.93	476.94
	DRT-8B	95.50	194.87	294.83	390.00	459.57	516.67	514.38
	DRT-14B	94.27	195.28	293.20	382.22	439.87	474.03	472.84
LitEval-Corpus	DRT-7B	95.97	193.87	293.37	395.45	492.04	831.38	913.88
	DRT-8B	91.12	189.58	290.99	393.68	486.73	876.08	1049.21
	DRT-14B	94.90	194.57	294.72	391.62	486.15	812.47	919.86
WMT24-Literary	DRT-7B	95.89	195.91	290.76	394.55	493.93	966.17	1870.02
	DRT-8B	90.71	196.56	268.99	353.15	458.29	902.59	1764.82
	DRT-14B	88.43	180.07	258.03	353.63	453.92	876.82	1727.59

Table 11: Actual thinking tokens of DRT models on literary translation tasks with varying thinking budgets.

Model	Original	No QS			QS		
		Budget=0	Budget=500	Budget=1000	Budget=0	Budget=500	Budget=1000
Qwen3-0.6B	69.293	69.756 (+0.463)	69.417 (+0.124)	69.445 (+0.152)	69.112 (-0.181)	69.538 (+0.246)	69.410 (+0.117)
Qwen3-1.7B	82.327	80.913 (-1.415)	82.896 (+0.568)	82.930 (+0.602)	81.438 (-0.889)	82.878 (+0.551)	83.275 (+0.947)
Qwen3-4B	88.695	88.623 (-0.072)	89.734 (+1.039)	89.888 (+1.193)	88.883 (+0.188)	89.725 (+1.030)	89.835 (+1.140)
Qwen3-8B	91.253	91.275 (+0.022)	91.618 (+0.365)	91.747 (+0.494)	91.173 (-0.080)	91.653 (+0.400)	91.657 (+0.404)
Qwen3-14B	92.199	92.241 (+0.042)	92.482 (+0.283)	92.529 (+0.330)	92.294 (+0.095)	92.448 (+0.249)	92.472 (+0.272)
Qwen3-32B	92.197	92.415 (+0.219)	92.502 (+0.305)	92.497 (+0.301)	92.365 (+0.169)	92.282 (+0.085)	92.334 (+0.138)

Table 12: Post-editing GRB score with and without QS at different budgets.

Model	Original	No QS			QS		
		Budget=0	Budget=500	Budget=1000	Budget=0	Budget=500	Budget=1000
Qwen3-0.6B	67.477	67.757 (+0.281)	67.936 (+0.459)	67.989 (+0.512)	67.987 (+0.510)	68.085 (+0.608)	68.057 (+0.580)
Qwen3-1.7B	81.727	79.159 (-2.568)	81.988 (+0.261)	82.305 (+0.578)	80.260 (-1.467)	82.445 (+0.718)	82.730 (+1.003)
Qwen3-4B	88.588	88.625 (+0.038)	89.936 (+1.349)	90.057 (+1.470)	88.755 (+0.168)	89.885 (+1.298)	90.031 (+1.444)
Qwen3-8B	91.208	90.440 (-0.768)	92.007 (+0.799)	92.018 (+0.810)	89.499 (-1.709)	92.051 (+0.843)	91.933 (+0.725)
Qwen3-14B	92.261	92.613 (+0.352)	92.763 (+0.501)	92.853 (+0.592)	92.642 (+0.381)	92.757 (+0.495)	92.810 (+0.549)
Qwen3-32B	92.256	92.767 (+0.511)	92.923 (+0.667)	92.864 (+0.608)	92.816 (+0.560)	92.768 (+0.512)	92.899 (+0.643)

Table 13: Post-editing GRF score with and without QS at different budgets.