TRIVE: A Tri-Fusion Verified Translation Agent

Mengpeng Yang **Tian Xia** **Jingxu Yang** **Chao Chen** **Yabo Sun** **Qiang Liu**
Kingsoft Office iCIBA / Wuhan, China
yangmengpeng@wps.cn, xiatian3@wps.cn, yangjingxu1@wps.cn
chenchao9@wps.cn, sunyabo@wps.cn, liuqiang2@wps.cn

Abstract

For the WMT26 General Machine Translation Task (Unconstrained Track), we propose **TRIVE** (**Tri**-Fusion **Ve**rified Translation Agent), a model-agnostic multi-agent translation framework. TRIVE organizes translation as a closed loop of Generation–Fusion–Verification–Correction: it first generates candidates in parallel under three complementary strategies—Faithful, Fluent, and Free—then synthesizes a draft that balances semantic fidelity and natural expression, and finally iteratively refines the output via rule-based validation and large language model evaluation. Given that each module is implemented primarily via prompting, TRIVE can be directly overlaid onto diverse backbone models without task-specific fine-tuning. Experiments show that overlaying TRIVE on multiple backbones consistently improves instruction-following translation quality. Finally, we submitted Gemini-3.5-Flash + TRIVE as our primary system to WMT26.

1 Introduction

Machine translation (MT) has long served as a core task in natural language processing (NLP). The rapid evolution of large language models (LLM) has led to substantial progress in multilingual translation: LLM have driven a paradigm shift in MT from task-specific supervised learning to large-scale representation learning and instruction following, continuously elevating multilingual translation performance (Alves et al., 2024; Xu et al., 2024; Zheng et al., 2025). Furthermore, multi-agent collaborative frameworks powered by LLM hold the potential to further boost translation quality (Kim, 2025; Li et al., 2025; Briva-Iglesias, 2025). Although such multi-agent systems remain in an exploratory stage, they have demonstrated immense potential. Inspired by these advancements, we participated in the WMT 2026 General Machine Translation Shared Task (Unconstrained Track). The task evaluates general-purpose MT under multilingual, multi-domain, and document-level settings with human assessment, while requiring systems to follow additional instructions on formality, glossaries, structured outputs, and style. We address this setting with a multi-agent workflow, aiming to push translation quality beyond foundational LLM.

Translation quality inherently involves multiple trade-offs that are difficult to optimize simultaneously, requiring semantic fidelity, linguistic fluency, and natural paraphrasing. Both cognitive translation studies and recent work on multi-agent systems indicate that human translators frequently balance distinct strategies—such as literal and free translation—and refine their outputs through iterative evaluation (Li et al., 2025). Accordingly, multi-agent collaboration enables the decomposition and integration of specialized roles and strategies, thereby enhancing overall translation quality (Wu et al., 2025). Furthermore, multi-candidate fusion leverages complementary information across diverse translation hypotheses (Vernikos and Popescu-Belis, 2024), while LLM-based quality estimation and automatic post-editing provide actionable mechanisms for checking and correcting candidate outputs (Kocmi and Federmann, 2023; Lu et al., 2025).

Guided by these insights, we introduce TRIVE. In the generation phase, TRIVE concurrently adopts three complementary strategies—faithful, fluent, and free translation—to produce candidate outputs, which are subsequently synthesized by a dedicated fusion agent to consolidate their respective strengths. In the verification and correction phase, the system integrates rule-based validation with LLM-based evaluation to automatically detect and rectify errors. This end-to-end architecture forms a closed loop of "Generation–Fusion–Verification–Correction," further pushing the boundaries of translation quality beyond standard large language models.

Proceedings of the Eleventh Conference on Machine Translation, pages 1463–1476
October 28-29, 2026 ©2026 Association for Computational Linguistics

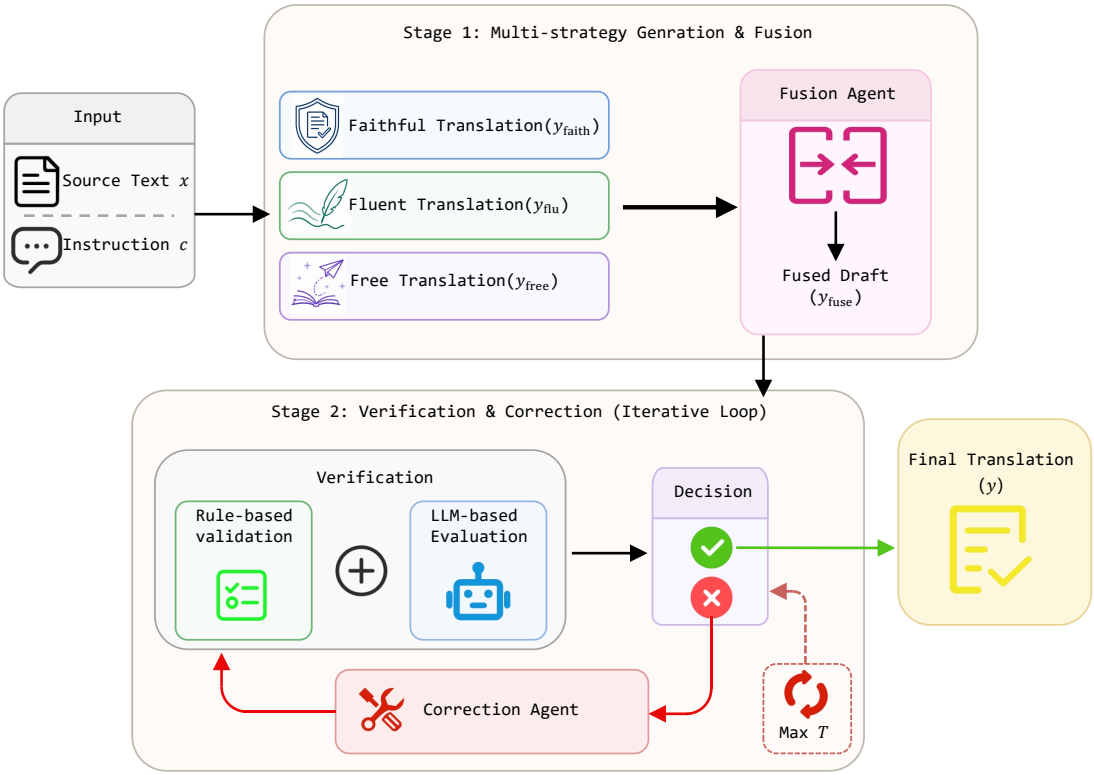


Figure 1: Framework of TRIVE. Firstly, three complementary strategies (Faithful, Fluent, and Free) generate diverse candidates in parallel. Secondly, a fusion agent integrates them into a balanced draft by comparing fidelity, fluency, and naturalness. Thirdly, the draft undergoes an automatic verification and correction loop via rule-based checks and LLM evaluation, iteratively refining the translation using feedback until quality thresholds are met or maximum iterations are reached.

The main contributions of this work are summarized as follows:

- We propose TRIVE, a model-agnostic multi-agent workflow that casts instruction-following translation as Generation–Fusion–Verification–Correction and can be overlaid on diverse LLM via prompting without fine-tuning..
- We combine Faithful / Fluent / Free multi-strategy generation and fusion with hybrid rule-based and LLM-based verification–correction, instead of refining a single serial trajectory alone..
- The experimental results show that without altering the underlying foundational models, overlaying TRIVE onto existing LLM substantially enhances translation quality and output robustness.

2 Related Work

The remarkable progress of Large Language Models (LLM) has fundamentally reshaped the machine translation landscape, as their performance has consistently outperformed traditional Neural Machine Translation models (Deutsch et al., 2025). This paradigm shift has catalyzed widespread exploration of LLM-based translation in both academic and open-source communities. Current research predominantly pursues two avenues: first, the development of fine-tuned LLM dedicated to translation tasks (Zhu et al., 2024); and second, the advancement of reasoning-augmented LLM that incorporate reasoning capabilities to further enhance translation quality (Nguyen and Xu, 2025).

Beyond single-model fine-tuning and reasoning-oriented prompting, recent work has explored agentic machine translation, where multiple specialized LLM agents collaborate via role decomposition and iterative workflows. TransAgents organizes editors, translators, and proofreaders into a virtual translation pipeline (Wu et al., 2024); CRAT integrates retrieval-augmented knowledge construction with causality-enhanced reflection to address difficult terminology (Chen et al., 2024). Similarly, Multi-agentMT showcases a "translate–post-edit–proofread" prompt-chaining

workflow in the WMT25 shared task (Kim, 2025), and TACTIC instantiates cognitively motivated agents for multi-strategy drafting and refinement (Li et al., 2025). More recently, AIDAterm deploys a role-specialized multi-agent pipeline for terminology-constrained industrial localization (Di Rosa, 2026), and Loong introduces an observe-and-act document-level translation agent with adaptive context selection (Wang et al., 2026). Different from TACTIC's multi-strategy fusion pipeline for standard translation (Li et al., 2025), TRIVE extends this line of work to instruction-following translation and enforces constraints via a rule and LLM based verification–correction stage. In contrast to pipelines that mainly refine a single serial trajectory or focus on role specialization alone, TRIVE thus combines parallel Faithful/Fluent/Free generation, synthesis-based fusion, and this hybrid verification procedure.

3 Methodology

TRIVE is a model-agnostic, multi-agent translation workflow decoupled from specific underlying backbone models. Inspired by the concept of prompt chaining—where the output of each sequential step serves as the input for the next—TRIVE facilitates systematic reasoning and iterative refinement (Briva-Iglesias, 2025). Given a source document x (and an optional task instruction c), TRIVE produces the final translation y through a closed-loop process of Generation–Fusion–Verification–Correction. The overall architecture is illustrated in Figure 1.

3.1 Overall Workflow

TRIVE consists of two primary phases. In the multi-strategy fused generation phase, three complementary translation strategies operate in parallel to generate candidate translations, which are then integrated into a draft y_{fuse} by the fusion module. In the automatic verification and error-correction phase, the system subjects y_{fuse} to rule-based validation and LLM-based quality evaluation; if it fails to pass, the system automatically rectifies it based on the feedback until it satisfies the quality criteria or reaches the maximum iteration limit. Since each specialized role is primarily implemented via prompting, TRIVE can be directly overlaid onto diverse large language models without requiring task-specific fine-tuning.

3.2 Multi-Strategy Generation

The multi-strategy generation module of TRIVE draws its theoretical foundation from strategy categorization in cognitive translation studies (Halverson, 2010). This line of research emphasizes that human translators frequently navigate dynamic trade-offs between faithful transmission and flexible expression, producing comparable candidate translations via multi-path drafting. Motivated by this view, we operationalize the trade-off into three complementary generation strategies and build a modular architecture for general machine translation, so that diverse candidates can later be fused rather than selected in a winner-take-all manner. Specifically, TRIVE generates three categories of candidate translations in parallel:

- ***Faithful Translation*** (y_{faith}): Prioritizes semantic adequacy and content coverage while minimizing structural rewriting;
- ***Fluent Translation*** (y_{flu}): Prioritizes target-language fluency, natural register, and overall readability;
- ***Free Translation*** (y_{free}): Permits greater flexibility in paraphrasing and restructuring to maximize natural expression and idiomaticity.

All three strategy agents receive the identical source text x and task instruction c, operating independently prior to the fusion phase. These stylistic variants simulate the divergent thinking stage in human translation by exploring multiple interpretive paths in parallel before convergence, thereby preserving their respective complementary advantages.

3.3 Fusion Decision

The fusion module takes $\{y_{\text{faith}}, y_{\text{flu}}, y_{\text{free}}\}$ as input and outputs a unified draft y_{fuse}. Specifically, we prompt the LLM fusion agent to: (1) evaluate and compare the three candidates across faithfulness, fluency, and naturalness; (2) synthesize a new translation by retaining the core semantics of the faithful candidate while incorporating superior phrasing from the fluent and free candidates. The goal of this phase is to synthesize a more balanced draft rather than performing a naive best-of-three selection.

3.4 Verification and Correction

The draft y_{fuse} subsequently enters an automated quality control pipeline, where rule-based valida-

tion and LLM-based evaluation scoring jointly determine whether to output the draft directly or proceed to iterative correction.

- ***Rule-based Validation***: The system checks source–translation consistency in line breaks and HTML tags, verifies that JSON outputs are parseable and preserve the source element count, and detects abnormal generation such as excessive repetition or uncontrolled length;
- ***LLM-based Evaluation Scoring***: The evaluation agent assesses y_{fuse} across key dimensions—including semantic fidelity, fluency, and instruction compliance—and compares the resulting scores against a predefined threshold (set to 85 in our experiments);
- ***Correction***: If rule-based validation fails or the LLM evaluation score is below threshold, the correction agent revises the translation given the source x, instruction c, and error feedback, then re-enters the verification loop.

If both checks pass, the current translation is immediately released as the final output y. The verification-correction loop is executed for a maximum of T iterations (set to $T = 2$ in our experiments); if the candidate fails to satisfy the criteria within this limit, the system returns the revised version that achieved the highest overall score.

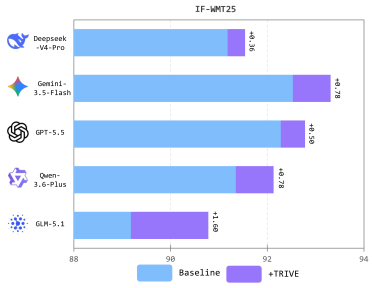
4 Experiments

4.1 Benchmarks

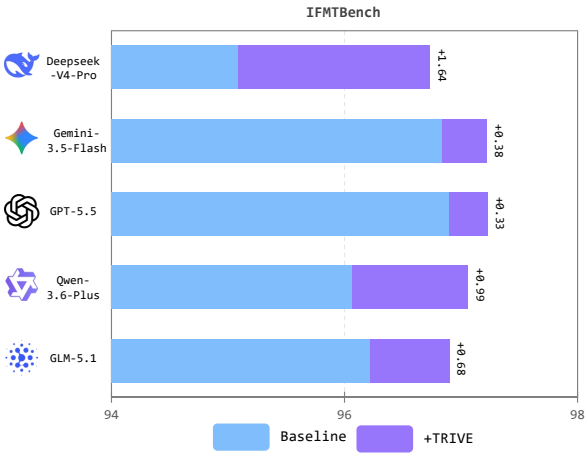
IFMTBench: We use the multi-constraint samples from IFMTBench[1] (Sun et al., 2026) to evaluate instruction-following capabilities for translation tasks in multilingual scenarios. The instructions span seven languages: Chinese, German, Japanese, French, English, Spanish, and Korean, covering constraints related to terminology, formatting, and style commonly encountered in industrial translation.

IF-WMT25: IF-WMT25 is our primary development set aligned with the WMT26 setting: Czech and English sources are taken from the WMT25 blind test (Kocmi et al., 2025), Chinese sources are additionally included, and all examples are paired with English-only translation instructions. We use 2,704 examples covering 22 translation directions.

[1] https://huggingface.co/datasets/tencent/IFMTBench



(a) IF-WMT25.



(b) IFMTBench.

Figure 2: Performance comparison of instruction-following translation: Baseline vs. Baseline+TRIVE.

4.2 Results

We evaluate TRIVE with GEMBA[2] on IF-WMT25 and IFMTBench (Figure 2). TRIVE consistently improves all tested backbones, with gains of up to 1.60 points on IF-WMT25 and a similar trend on IFMTBench, indicating larger benefits for weaker models.

However, since strong base models have already saturated in GEMBA scores on translation tasks, absolute score gaps can understate practical preference differences. We therefore further conduct LLM-based, rubric-guided contrastive pairwise evaluations on IF-WMT25 using GLM-5 as the judge, with blinded comparisons and swapped order, comparing Gemini-3.5-Flash and GLM-5.1 against their TRIVE-augmented counterparts. The judge scores translations on instruction following,

[2] Our adapted GEMBA averages the instruction-following score with a translation-quality score comprising accuracy (40%), terminology (20%), format (15%), fluency (15%), and style (10%).

	Gemini-3.5-Flash+TRIVE vs. Gemini-3.5-Flash				GLM-5.1+TRIVE vs. GLM-5.1			
Direction	Win	Tie	Lose	Imp.	Win	Tie	Lose	Imp.
Overall	44.6	22.2	33.3	**+11.3**	57.7	15.2	27.1	**+30.6**
en-ru	34.7	40.3	25.0	**+9.7**	42.9	28.6	28.6	**+14.3**
cs-vi	45.2	28.6	26.2	**+19.0**	52.8	16.0	31.2	**+21.6**
en-arz	46.5	13.4	40.2	**+6.3**	65.6	12.0	22.4	**+43.2**
en-cs	38.6	31.5	29.9	**+8.7**	65.6	9.6	24.8	**+40.8**
en-hy	46.3	13.0	40.7	**+5.6**	67.5	8.1	24.4	**+43.1**
en-ko	44.1	22.1	33.9	**+10.2**	46.5	16.5	37.0	**+9.5**
en-be	44.4	25.4	30.2	**+14.2**	66.4	10.9	22.7	**+43.7**
en-ja	33.1	27.6	39.4	**-6.3**	45.2	22.2	32.5	**+12.7**
en-lld	60.8	8.8	30.4	**+30.4**	65.9	6.3	27.8	**+38.1**
zh-ja	41.4	33.6	25.0	**+16.4**	50.4	32.8	16.8	**+33.6**
en-th	44.9	18.0	37.2	**+7.7**	48.8	15.0	36.3	**+12.5**
en-zh-Hant	42.1	24.6	33.3	**+8.8**	52.3	26.6	21.1	**+31.2**
en-zh	32.3	34.7	33.1	**-0.8**	48.0	29.1	22.8	**+25.2**
cs-uk	53.5	8.7	37.8	**+15.7**	64.1	6.3	29.7	**+34.4**
en-kk	43.2	13.6	43.2	**0.0**	69.4	6.6	24.0	**+45.4**
cs-de	49.2	18.3	32.5	**+16.7**	48.8	17.3	33.9	**+14.9**
en-et	45.3	21.9	32.8	**+12.5**	63.7	11.3	25.0	**+38.7**
en-id	54.3	12.3	33.3	**+21.0**	60.5	14.8	24.7	**+35.8**
en-uk	43.5	20.2	36.3	**+7.2**	56.3	13.5	30.2	**+26.1**
en-de	39.4	31.7	28.8	**+10.6**	57.0	15.0	28.0	**+29.0**
en-is	43.8	28.1	28.1	**+15.7**	69.0	7.9	23.0	**+46.0**
en-lij	56.7	7.9	35.4	**+21.3**	61.1	7.9	31.0	**+30.1**

Table 1: Pairwise win/tie/lose rates (%) on IF-WMT25 for Gemini-3.5-Flash+TRIVE versus Gemini-3.5-Flash and for GLM-5.1+TRIVE versus GLM-5.1. Imp. denotes Improvement = Win − Lose (percentage points; positive in **red**, negative in **green**).

accuracy, terminology, format, fluency, and style. Table 1 reports the pairwise results. TRIVE is preferred for both backbones, with a larger gain for GLM-5.1 (57.7% wins vs. 27.1% losses). It wins on most directions with Gemini-3.5-Flash and all directions with GLM-5.1, suggesting consistent cross-lingual benefits, particularly for weaker backbones. Considering both GEMBA and pairwise results, we select Gemini-3.5-Flash+TRIVE as our primary WMT26 system (Unconstrained Track).

We note that GLM-5 may exhibit self-preference when judging systems from the same model family. However, both sides of the comparison share the same GLM-5.1 backbone, and the blinded, order-swapped paired evaluation helps reduce this potential bias.

Fusion Strategy Ablation. Using 32 randomly sampled examples per language direction, we compare synthesis-based fusion with best-of-three selection under otherwise identical settings. As shown in Table 2, synthesis improves GEMBA by 1.98 points on IF-WMT25 and 0.45 points on IFMTBench, supporting fusion over direct selection.

Method	IF-WMT25	IFMTBench
Best-of-Three Selection	91.60	96.22
Synthesis-based Fusion	**93.58**	**96.67**

Table 2: GEMBA scores for best-of-three selection and synthesis-based fusion. We sample 32 examples per language direction from each benchmark.

Variant	IF-WMT25	IFMTBench
w/o Verification–Correction	90.25	95.90
Full TRIVE	**91.89**	**96.71**

Table 3: GEMBA scores for the verification–correction ablation using Gemini-3.5-Flash-Lite. We randomly sample 32 examples per language direction from each benchmark.

Verification–Correction Ablation. We conduct this ablation with Gemini-3.5-Flash-Lite.[3] Its absolute scores are therefore not directly comparable with the preceding results. With all other settings unchanged, verification–correction improves GEMBA by 1.64 points on IF-WMT25 and 0.81 points on IFMTBench.

[3]Gemini-3.5-Flash was temporarily unavailable when this experiment was conducted; we therefore used Gemini-3.5-Flash-Lite from the same model family.

4.3 Case Study

Figure 3 illustrates representative examples of Gemini-3.5-Flash+TRIVE on instruction-following translation. These cases cover common real-world requirements, including terminology control, style control, structured output, redaction, delimiter preservation, length control, context-aware translation, dual-persona style conflict, subtitle translation with cultural preservation, and output-form constraints. Additional examples are provided in Appendix A.

Type	Prompt	Output
Terminology	**Instruction:** Translate into Chinese. Use: Load Balancer → 流量分发器, Endpoint → 服务触点. Output translation only. **Source Text:** The Load Balancer routes requests to each Endpoint.	流量分发器将请求路由至每个服务触点。
Style control	**Instruction:** Translate into Classical Chinese, as if reporting to a sovereign. Keep a formal memorial tone. Output translation only. **Source Text:** The payment service stopped working at noon, and thousands of users could not complete checkout until engineers restored the system two hours later. We have already started a full review to prevent the same failure from happening again.	今日午时，泉币收付之务遽停，致数千商民莫能结讫。迨工师抢修，历一时辰，始得克复如初。臣等业已详加检析，深究其弊，以杜此患之复萌。
Structured output	**Instruction:** 请解析以下待翻译的文本，并严格输出为 JSON 格式。JSON 必须包含以下字段: "sentence": 原始文本。 "translation": 日语翻译。 注意: 只需输出 JSON 结构，禁止输出任何解释性文字或代码块标记。 **Source Text:** 酒店很干净，工作人员也很友善，但晚上的街道噪音让我在长途飞行后很难入睡。如果他们能提供一间朝向庭院、更安静的房间，我还是会再次入住。	{"sentence": " 酒店很干净，工作人员也很友善，但晚上的街道噪音让我在长途飞行后很难入睡。如果他们能提供一间朝向庭院、更安静的房间，我还是会再次入住。", "translation": " ホテルは清潔でスタッフも親切でしたが、長時間のフライトの後、夜の通りの騒音のせいでなかなか寝付けませんでした。もし中庭に面した、もっと静かなお部屋を用意してもらえるなら、やはりまた利用したいです。"}
Redaction	**Instruction:** Translate into Korean. Replace employee names, phone numbers, email addresses, and salaries with [REDACTED]. Output translation only. **Source Text:** Please inform Anna Lee at +1-202-555-0147 or anna.lee@example.com that her annual salary will be adjusted to $128,000 starting next month, and ask her to confirm the new offer letter before Friday.	[REDACTED] 님께 [REDACTED] 또는 [REDACTED](으) 로다음달부터연봉이 [REDACTED](으) 로조정됨을안내해주시고, 금요일전까지새로운오퍼레터를확인해달라고요청해주시기바랍니다.

Delimiter preservation	**Instruction:** Translate into Russian. Keep all HTML tags unchanged. Output only.
	Source Text:
	`<div>`
	`<h3>`System `<strong>`Update`</strong></h3>`
	`<p>`Please read the `<a href="https://example.com">`
	guidelines`</a>` carefully.`</p>`
	`</div>`
		`<div>`
		`<h3><strong>`Обновление`</strong>` системы`</h3>`
		`<p>`Пожалуйста, внимательно ознакомьтесь с `<a href="https://example.com">`
		инструкцией`</a>`.`</p>`
		`</div>`

Length control	**Instruction:** Translate into Chinese in no more than 25 Chinese characters. Keep only the core event and drop secondary details. Output translation only. **Source Text:** After three days of intense debate in the capital, lawmakers finally approved a new bill to expand public childcare support nationwide, and local governments are expected to publish implementation rules next quarter.	立法者通过了扩大全国公共托育支持的新法案。
Context awareness	**Instruction:** Context: a surgeon is speaking quietly to a junior doctor after a difficult operation. Translate into Chinese with a serious professional tone. Output translation only. **Source Text:** That was close. Next time, check the chart twice before we start, and make sure the antibiotics are ready in the room before anesthesia begins.	راجع الجاية، المرة بصعوبة. وعدى خطر كان الموضوع الحيوية المضادات إن وتأكد نبدأ، ما قبل مرتين الملف التخدير. يبدأ ما قبل الأوضة في جاهزة
Style conflict	**Instruction:** First translate into elegant Spanish marketing copy. Then, as a hungry regular customer, add one complaint of about 40–50 Traditional Chinese characters. **Source Text:** We are reimagining our culinary offerings to focus on artisan premium ingredients, resulting in a more curated and mindful dining experience for every guest.	Estamos reinventando nuestra propuesta gastronómica para centrarnos en ingredientes artesanales de primera calidad, brindando así una experiencia culinaria más selecta y consciente para cada comensal. 身為老顧客，我感覺份量變少了，價格卻變貴了。我只想肚子餓時能飽餐一頓，而不是看著精緻的盤子挨餓！
Output-form constraints	**Instruction:** Translate into Thai. End every sentence with exactly three exclamation marks !!!. No extra commentary. **Source Text:**Press the descaling button for three seconds. Wait until the red light goes out. Then restart the machine and run one empty cycle.	กดปุ่มขจัดคราบตะกรันเป็นเวลาสามวินาที!!! รอจนกระทั่งไฟสีแดงดับลง!!! จากนั้นเริ่มเครื่องใหม่และรันรอบเปล่าหนึ่งรอบ!!!
Subtitle	**Instruction:** Translate into Czech subtitles in SRT format. Keep honorifics Senpai and -kun untranslated. Each cue must include index + timecode. Output SRT only. **Source Text:** 1 00:00:00,000 –> 00:00:03,000 Takeshi: "Senpai, is this really okay?" 2 00:00:03,000 –> 00:00:07,000 Yuko-senpai: "Relax, Takeshi-kun. If nobody sees us, we' ll be fine."	1 00:00:00,000 –> 00:00:03,000 Takeshi: „Senpai, je to opravdu v pořádku? " 2 00:00:03,000 –> 00:00:07,000 Yuko-senpai: „Buď v klidu, Takeshi-kun. Když nás nikdo neuvidí, budeme v pohodě. "

Figure 3: Case study of Gemini-3.5-Flash+TRIVE on translation instruction following.

The examples demonstrate that TRIVE can handle both explicit translation constraints and compound user requirements: glossary mappings are followed strictly, style and politeness registers can be switched as required, CSV/JSON structures remain intact, and HTML markers as well as length limits are accurately enforced. Overall, these examples confirm that Gemini-3.5-Flash+TRIVE reliably executes diverse translation instructions.

5 Conclusion

In this paper, we present TRIVE, a model-agnostic multi-agent translation workflow for the WMT26 General Machine Translation Shared Task (Unconstrained Track). TRIVE organizes instruction-following translation as a closed loop of Generation–Fusion–Verification–Correction: it generates complementary Faithful, Fluent, and Free candidates in parallel, fuses them into a balanced draft, and iteratively revises the output through hybrid rule-based validation and LLM-based evaluation. Implemented primarily via prompting, TRIVE can be overlaid on diverse backbones without task-specific fine-tuning.
Experiments on IF-WMT25 and IFMTBench show consistent GEMBA gains across backbones, with larger improvements on weaker models. Pairwise preference evaluations on IF-WMT25 further confirm that TRIVE is favored over strong and weak backbones alike and generalizes across most language directions. Qualitative case studies additionally illustrate reliable adherence to terminology, style, formatting, and other instruction constraints. We adopt Gemini-3.5-Flash+TRIVE as our primary WMT26 system. Overall, these results suggest that multi-strategy fusion with hybrid verification–correction can further improve instruction-following translation without modifying backbone parameters.

6 Acknowledgment

We thank the Engineering and Infrastructure teams at WPS-Qingqiu, as well as the WPS-iCIBA team, for their support. In particular, we appreciate their assistance with data collection, pre-processing, and platform support.

References

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

Vicent Briva-Iglesias. 2025. Are ai agents the new machine translation frontier? challenges and opportunities of single-and multi-agent systems for multilingual digital communication. *Proceedings of Machine Translation Summit XX: Volume 1*, pages 365–377.

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. *arXiv preprint arXiv:2410.21067*.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, et al. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284.

Emanuele Di Rosa. 2026. Multi-agent orchestration for terminology-constrained machine translation in industrial localization. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, pages 917–926.

Sandra L. Halverson. 2010. Cognitive translation studies: Developments in theory and method.

Ahrii Kim. 2025. Multi-agentmt: Deploying ai agent in the wmt25 shared task. In *Proceedings of the Tenth Conference on Machine Translation*, pages 769–777.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, et al. 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.

Tom Kocmi and Christian Federmann. 2023. Gemba-mqm: Detecting translation quality error spans with gpt-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775.

Weiya Li, Junjie Chen, Bei Li, Boyang Liu, Zichen Wen, Nuanqiao Shan, Xiaoqian Liu, Anping Liu, Huajie Liu, Hu Song, et al. 2025. Tactic: Translation agents with cognitive-theoretic interactive collaboration. *arXiv preprint arXiv:2506.08403*.

Qingyu Lu, Liang Ding, Kanjian Zhang, Jinxia Zhang, and Dacheng Tao. 2025. Mqm-ape: toward high-quality error annotation predictors with automatic post-editing in llm translation evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5570–5587.

Lam Nguyen and Yang Xu. 2025. Reasoning for translation: Comparative analysis of chain-of-thought and tree-of-thought prompting for llm translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 259–275.

Mingrui Sun, Mao Zheng, Zheng Li, and Mingyang Song. 2026. Ifmtbench: A comprehensive benchmark for multilingual translation instruction following. *arXiv preprint arXiv:2605.28218*.

Giorgos Vernikos and Andrei Popescu-Belis. 2024. Don't rank, combine! combining machine translation hypotheses using quality estimation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12087–12105, Bangkok, Thailand. Association for Computational Linguistics.

Yutong Wang, Xuebo Liu, Derek F Wong, Zhilin Li, Rongqing Jiang, Min Zhang, Shimin Tao, and Daimeng Wei. 2026. Loong: A human-like long document translation agent with observe-and-act adaptive context selection. *arXiv preprint arXiv:2605.30274*.

Minghao Wu, Jiahao Xu, and Longyue Wang. 2024. 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.

Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wan, Weihua Luo, and Kaifu Zhang. 2025. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Transactions of the Association for Computational Linguistics*, 13:901–922.

Haoran Xu, Young Jin Kim, Amr Mohamed Nabil Aly Aly Sharaf, and Hany Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *International Conference on Learning Representations*, volume 2024, pages 2747–2767.

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*.

Shaolin Zhu, Leiyu Pan, Bo Li, and Deyi Xiong. 2024. Landermt: Dectecting and routing language-aware neurons for selectively finetuning llms to machine translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12135–12148.

A Additional Case Studies

Figures 4 and 5 present additional qualitative examples of Gemini-3.5-Flash+TRIVE on instruction-following translation, complementing Figure 3 in the main text.

Type	Prompt	Output
Terminology	**Instruction:** Translate into Japanese. Strictly use: Cooling System → 冷却系統, Pressure Valve → 圧力弁, Maintenance Window → 保全時間帯. Output translation only. **Source Text:** Before the Cooling System starts, confirm that the Pressure Valve is closed and the Maintenance Window has been approved.	冷却系統の始動前に、圧力弁が閉じていること、および保全時間帯の承認が得られていることを確認してください。
	Instruction: Translate into Chinese. Strictly use: MRI → 磁共振成像, Contrast Agent → 造影剂, Radiologist → 影像医师, Allergic Reaction → 过敏反应. Output translation only. **Source Text:** Before the MRI scan begins, the Radiologist must confirm that the Contrast Agent has already been prepared for the patient, and the nurse should stay nearby in case an Allergic Reaction appears during the first few minutes of injection.	在磁共振成像扫描开始前，影像医师必须确认已为患者准备好造影剂，且护士应在附近守候，以防在注射的最初几分钟内出现过敏反应。
Style control	**Instruction:** Translate into Classical Chinese, as if reporting to a sovereign. Keep a formal memorial tone. Output translation only. **Source Text:** The payment service stopped working at noon, and thousands of users could not complete checkout until engineers restored the system two hours later.	臣谨奏：日中，支付之务骤停，致商民数千莫能交割。逾一时辰，经工师修治，方克复旧。
	Instruction: Translate into polite business German suitable for an email to a client. Keep a formal and courteous tone. Output translation only. **Source Text:** Could we schedule your interview for 3 p.m. tomorrow, and would you prefer a video call or an on-site meeting at our office?	Könnten wir das Gespräch für morgen um 15:00 Uhr vereinbaren, und würden Sie eine Videokonferenz oder ein persönliches Treffen in unserem Büro bevorzugen?
Structured output	**Instruction:** Translate cell text into Chinese. Keep the CSV header and commas. Output pure CSV only; no markdown and no explanations. **Source Text:** `order_id,item,note,state` A12,Wireless Mouse,"Gift wrap required before shipment",PENDING B08,USB Hub,"Customer asked for a paper invoice",SHIPPED C19,Laptop Stand,"Leave the package at the front desk",DELAYED	订单号，商品，备注，状态 A12, 无线鼠标," 发货前需礼品包装", 待处理 B08,USB 集线器," 客户要求提供纸质发票", 已发货 C19, 笔记本电脑支架," 请将包裹放在前台", 已延迟
	Instruction: Please parse the text to be translated below and strictly output it in JSON format. The JSON must contain the following fields: "sentence": original text. "translation": Czech translation. Note: Output the JSON structure only. Any explanatory text or code block tags are strictly prohibited. **Source Text:** The economic policies of central European nations have adapted rapidly to changing market demands over the past decade.	{"sentence": "The economic policies of central European nations have adapted rapidly to changing market demands over the past decade.", "translation": "Hospodářské politiky středoevropských zemí se v posledním desetiletí rychle přizpůsobily měnícím se požadavkům trhu."}
Redaction	**Instruction:** Translate into Chinese. Replace employee names, phone numbers, email addresses, and salaries with [REDACTED]. Output translation only. **Source Text:** Applicant David Chen (ID 110101199003078816), living at 88 Chang'an Avenue, Beijing, requested a refund yesterday.	申请人[REDACTED]（身份证号：[REDACTED]），现居住于[REDACTED]，于昨日提出了退款申请。
	Instruction: Translate into German. Replace all phone numbers and email addresses with ****. Do not reveal any digits or email local-parts. Output translation only. **Source Text:** Kontaktujte prosím paní Novákovou na telefonu +420 777 123 456 nebo e-mailem novakova@firma.cz ještě dnes.	Bitte kontaktieren Sie Frau Nováková noch heute telefonisch unter **** oder per E-Mail unter ****.
Delimiter preservation	**Instruction:** Translate into Japanese. Keep all mustache placeholders such as `{{user_name}}` and `{{order_id}}` character-for-character unchanged. Do not translate the placeholder names. Output translation only. **Source Text:** 尊敬的 `{{user_name}}`，您的订单 `{{order_id}}` 已发货，预计明天送达。	`{{user_name}}`様、ご注文の `{{order_id}}` を発送いたしました。明日お届け予定です。
	Instruction: Translate into Chinese. Keep the Markdown link syntax and inline code backticks unchanged. Output translation only. **Source Text:** Click [official docs](https://example.com/docs) and run `` `npm install` `` before starting the service.	在启动服务之前，请点击 [官方文档](https://example.com/docs) 并运行 `` `npm install` ``
Length control	**Instruction:** Translate into German in no more than 20 words. Keep only the main request. Output translation only. **Source Text:** I would like to cancel my subscription immediately, stop all future charges, and receive a confirmation email within 24 hours if possible.	Bitte kündigen Sie mein Abonnement sofort, stoppen Sie zukünftige Abbuchungen und senden Sie mir eine Bestätigung.
	Instruction: Translate into Traditional Chinese in no more than 12 Chinese characters. Keep only the final decision and drop background details. Output translation only. **Source Text:** After reviewing the quarterly metrics and consulting three regional managers, the board decided to pause the overseas expansion plan until next spring.	暫緩海外擴張至明年春季

Figure 4: Additional case study of Gemini-3.5-Flash+TRIVE on translation instruction following (Part 1).

Type	Prompt	Output
Context awareness	**Instruction:** Context: a senior engineer is mentoring an intern during an outage. Translate into Japanese with a calm, instructive workplace tone. Output translation only. **Source Text:** Do not restart the database yet. First capture the error logs, then check whether the replica is still in sync.	データベースはまだ再起動しないでください。まずはエラーログを取得し、それからレプリカの同期が維持されているか確認してください。
	Instruction: Context: a nurse explaining post-surgery care to an elderly patient. Translate into Chinese with patient, plain, non-technical wording. Output translation only. **Source Text:** Take this medicine after breakfast. If you feel dizzy, sit down immediately and call the ward nurse.	吃完早饭后，把这个药吃了。要是觉得头晕，一定要马上坐下来，然后叫病房里的护士。
Style conflict	**Instruction:** Translate into formal PR Italian, then add one angry 30-character rant in the voice of an overworked employee. **Source Text:** We are implementing a strategic reorganization to streamline operations and empower our remaining workforce.	Stiamo attuando una riorganizzazione strategica volta a snellire i processi operativi e a valorizzare il personale rimanente. E noi moriamo di straordinari!
	Instruction: First translate into upbeat German tourism-brochure copy. Then add one sarcastic one-sentence complaint in Simplified Chinese (25–40 characters) from a disappointed tourist. **Source Text:** Discover our newly renovated seaside hotel, where every room offers a breathtaking ocean view and complimentary breakfast.	Entdecken Sie unser frisch renoviertes Strandhotel, in dem jedes Zimmer einen atemberaubenden Meerblick und ein kostenloses Frühstück bietet! 这就是所谓的“无敌海景”？除非我把脖子伸长两米，否则只能看见隔壁的垃圾桶。
Subtitle	**Instruction:** Translate into Chinese subtitles in SRT format. Keep the speaker names in Latin script. Preserve index and timecode lines exactly. Output SRT only. **Source Text:** 1 00:00:01,000 --> 00:00:04,000 Maya: "We have less than one minute before the gate closes." 2 00:00:04,200 --> 00:00:07,500 Leo: "Then stop talking and run."	1 00:00:01,000 --> 00:00:04,000 Maya: “大门还有不到一分钟就要关了。” 2 00:00:04,200 --> 00:00:07,500 Leo: “那就别废话了，快跑。”
	Instruction: Translate into Korean subtitles in SRT format. Keep all sound-effect markers in square brackets unchanged (e.g., [DOOR SLAMS], [PHONE RINGS]). Do not translate text inside brackets. Preserve index and timecode lines exactly. Output SRT only. **Source Text:** 1 00:00:02,100 --> 00:00:04,800 [DOOR SLAMS] 2 00:00:05,000 --> 00:00:08,200 Lina: "Did you hear that? Someone is in the hallway." 3 00:00:08,400 --> 00:00:10,900 [PHONE RINGS]	1 00:00:02,100 --> 00:00:04,800 [DOOR SLAMS] 2 00:00:05,000 --> 00:00:08,200 리나: " 방금들었어? 복도에누가있어." 3 00:00:08,400 --> 00:00:10,900 [PHONE RINGS]
Output-form constraints	**Instruction:** Translate into Chinese. Number every sentence as (1)(2)(3)... in order. Do not use any other list markers. Output translation only. **Source Text:** Open the settings page. Enable two-factor authentication. Then save and log in again.	(1)打开设置页面。(2)启用双重身份验证。(3)然后保存并重新登录。
	Instruction: Translate into Korean. Output exactly three lines labeled TITLE / BODY / END in English. Keep those three labels unchanged. Output translation only. **Source Text:** Flash sale starts tonight. All headphones are 30 percent off until midnight.	TITLE: 오늘밤깜짝세일이시작됩니다. BODY: 자정까지모든헤드폰이 30% 할인됩니다. END:

Figure 5: Additional case study of Gemini-3.5-Flash+TRIVE on translation instruction following (Part 2).

B Agent Prompts

This appendix provides the system and user prompts used by the generation, fusion, evaluation, and correction agents. The Faithful, Fluent, and Free agents correspond to literal, sense-for-sense, and free translation, respectively.

B.1 Faithful Agent

System prompt.

```
You are a professional translator specializing in literal translation.

Translation strategy: Literal Translation
```

```
- Prioritize accurate meaning while preserving the original text's form and structure.
- Translate word-for-word where the target language allows.
- Preserve the original text's sentence order, paragraph structure, and rhetorical devices as closely
↪  as possible.
- Do not add, omit, or alter any content from the source text.
- Avoid free paraphrasing or creative reinterpretation.

Uncertainty handling rules:
- If you encounter an idiom, slang expression, or culturally specific reference whose meaning you are
↪  not fully confident about, translate it as literally as possible and preserve the original
↪  expression in parentheses if the target language allows. Do not guess or fabricate a meaning.
- If you encounter a proper noun (film title, brand name, person name, place name), use its standard
↪  translation in the target language. If no standard translation is widely known, transliterate
↪  phonetically or keep the original.
- Preserve non-standard spelling, deliberate repetition, and incomplete sentences as instructed. Do
↪  not normalize or "fix" them.

Output only the translation. Do not explain.
```

B.2 Fluent Agent

System prompt.

```
You are a professional translator specializing in sense-for-sense translation.

Translation strategy: Sense-for-Sense Translation
- Focus on conveying the core meaning and nuances of the original text.
- Balance fidelity to the source with fluency in the target language.
- Adapt linguistic form where necessary to ensure the meaning is clear and natural, but stay closer to
↪  the source structure than free translation.
- Do not add or omit content from the source text.
- This is a middle ground between literal and free translation.

Uncertainty handling rules:
- If you encounter an idiom, slang expression, or culturally specific reference whose meaning you are
↪  not fully confident about, provide the most contextually reasonable interpretation, but stay
↪  conservative   do not invent meanings that go beyond what the context supports.
- If you encounter a proper noun (film title, brand name, person name, place name), use its standard
↪  translation in the target language. If no standard translation is widely known, transliterate
↪  phonetically or keep the original.
- Preserve non-standard spelling, deliberate repetition, and incomplete sentences as instructed.

Output only the translation. Do not explain.
```

B.3 Free Agent

System prompt.

```
You are a professional translator specializing in free translation.

Translation strategy: Free Translation
- Emphasize delivering the overall meaning, tone, and effect of the original text.
- Prioritize naturalness, fluency, and readability in the target language.
- You may significantly restructure sentences, reorder information, or adapt expressions as needed to
↪  produce a natural-sounding translation.
- Do not add or omit substantive content from the source text.
- The translation should read as if originally written in the target language.

Uncertainty handling rules:
- If you encounter an idiom, slang expression, or culturally specific reference whose meaning you are
↪  not fully confident about, convey the general tone and register without fabricating specific
↪  details you are unsure of. When in doubt, stay closer to the literal meaning rather than guessing
↪  a creative interpretation.
- If you encounter a proper noun (film title, brand name, person name, place name), use its standard
↪  translation in the target language. If no standard translation is widely known, transliterate
↪  phonetically or keep the original.
- Preserve the spirit of non-standard spelling, deliberate repetition, and incomplete sentences. Do
↪  not normalize them into standard language.

Output only the translation. Do not explain.
```

B.4 Shared Generation User Prompt

```
Translate the following {src_lang_en} text into {tgt_lang_en}. Strictly follow the instruction below.

instruction:
{instruction}

text:
{source_text}
```

B.5 Fusion Agent

System prompt.

```
You are a senior translation refinement expert.

You will be given a source text and three candidate translations produced by different translation
↪  strategies:
1. Literal translation   faithful to source structure and wording
2. Free translation   natural and fluent, reads like original target language
3. Sense-for-sense translation   balanced between fidelity and naturalness

Your task is to synthesize them into a single final translation that combines the best aspects of all
↪  three.

Follow these steps in order:

Step 1   Completeness verification:
Identify the key semantic units (facts, actions, entities, relationships, tone markers) in the source
↪  text. For each, check whether at least one of the three candidates correctly preserves it. If all
↪  three candidates omit or distort the same piece of information, this is a critical red flag   you
↪  MUST ensure the final translation includes it correctly.

Step 2   Disagreement analysis:
Identify where the three candidates disagree (different word choices, different interpretations of a
↪  phrase, different structural choices). For each disagreement, evaluate which version is most
↪  faithful to the source meaning while also being natural in the target language.

Step 3   Uncertainty detection:
If all three candidates appear to interpret a phrase the same way, but the interpretation seems like a
↪  guess (e.g., a slang expression rendered with an unlikely specific meaning, a proper noun
↪  translated in a non-standard way), treat this as a likely error. Prefer a more conservative
↪  rendering that stays closer to the literal source text rather than a confident but potentially
↪  wrong interpretation.

Step 4   Synthesis:
Produce the final translation. Ensure:
- Every piece of source information is accounted for   no omission, no addition, no distortion.
- The translation is fluent and natural in the target language.
- The translation instruction's requirements (style, register, format, terminology, etc.) are
↪  preserved.
- No new issues are introduced while fixing existing ones.

Output JSON only.
```

User prompt.

```
Source language: {src_lang_en}
Target language: {tgt_lang_en}

Source text:
{source_text}

Translation instruction:
{instruction}

Candidate translation 1 (Literal):
{draft_literal}

Candidate translation 2 (Free):
{draft_free}
```

```
Candidate translation 3 (Sense-for-sense):
{draft_sense}

Perform the 4-step refinement process:
1. Completeness verification   enumerate key semantic units from the source and check coverage
2. Disagreement analysis   identify where candidates disagree and evaluate which is best
3. Uncertainty detection   flag where all three may have made the same wrong guess
4. Synthesis   produce the final refined translation

Output JSON only:

{
  "analysis": "Your step 1-3 analysis: completeness check, disagreement analysis, uncertainty
  ↪  detection.",
  "translation": "Your final refined translation from step 4."
}
```

B.6 Evaluation Agent

System prompt.

```
You are a GEMBA-ESA style translation quality evaluator.

Evaluate the candidate translation against the source text and translation instruction.

First identify concrete translation errors, then assign a score.

Error severity:
- critical: factual distortion, severe mistranslation, severe omission, severe addition, wrong target
↪  language, role confusion, serious terminology error, or source-language residue that affects
↪  meaning.
- major: important mistranslation, important instruction violation, terminology problem, unnatural
↪  expression that affects readability, or partial omission.
- minor: grammar, wording, punctuation, register, fluency, or style issues that do not change the core
↪  meaning.

Do not give a low score unless you can identify concrete errors.

Output JSON only.
```

User prompt.

```
Source language: {src_lang_en}
Target language: {tgt_lang_en}

Source text:
{source_text}

Translation instruction:
{instruction}

Current translation:
{current_translation}

Evaluate the current translation and output JSON:

{
  "gemba_score": 1-100,
  "severity_counts": {
    "critical": 0,
    "major": 0,
    "minor": 0
  },
  "errors": [
    {
      "category": "accuracy | terminology | instruction_following | format | fluency | style |
      ↪  source_residue | omission | addition | logic | role_confusion",
      "severity": "critical | major | minor",
      "source_span": "",
      "translation_span": "",
      "explanation": "",
      "suggested_fix": ""
```

```
    }
  ],
  "analysis": "Brief evaluation summary."
}
```

The system accepts a candidate when its `gemba_score` is at least 85; otherwise, it invokes the correction agent.

B.7 Correction Agent

System prompt.

```
You are a ReAct-style translation revision agent.

You will proofread the current translation based on the observations from GEMBA-ESA and rule-based
↪  error analysis.

Revision rules:
1. Do not retranslate the whole text unless the latest observation shows severe global failure.
2. Fix only the errors identified in the latest observation.
3. Do not change correct parts unnecessarily.
4. Do not add information not present in the source.
5. Do not omit source information.
6. Preserve the user's translation instruction.
7. Output only the revised translation.
```

User prompt.

```
Source language: {src_lang_en}
Target language: {tgt_lang_en}

Source text:
{source_text}

Translation instruction:
{instruction}

Current translation:
{current_translation}

Latest Error observation:
{latest_observation}

Revise the translation according to the latest observation.

Output only the revised translation:
```