TARL: A Terminology-Aware Agentic Repair Loop for Language-Agnostic Translation

Shivam Mhaskar[*1] **Vrunda N. Sukhadia**[*1] **Anurag Deshmukh**[*1], **Manan Sharma**[*1], **Pawan Rajpoot**[1]
[1]Amazon
[*]Equal contribution

Correspondence: shivmhas@amazon.com, vrundans@amazon.com

Abstract

We describe **TARL** (Terminology-Aware Agentic Repair Loop), our submission to the WMT26 Terminology Translation Task, which targets document-level, terminology-constrained translation for morphologically rich and lower-resource directions (Spanish→Basque, English→Polish) and Traditional Chinese→English. Our system performs no fine-tuning: it is a sequential pipeline of four role-specialized LLM agents (translate, verify, review, and post-edit) in which a required target term, once inserted, is treated as locked so that later agents improve the translation around the terminology rather than altering it. Relevant glossary terms are retrieved per sentence by an exact, fuzzy n-gram matcher (with substring matching for CJK source), and for the sample-only Track 2 we first induce a glossary by extracting bilingual term pairs from the provided parallel data. The agents' instructions are optimized automatically with GEPA (Genetic-Pareto), an automated prompt optimization framework that uses natural language reflection and multi-objective Pareto evolutionary search to tune LLM prompts, rather than being hand-written. On the official WMT26 Track 1 evaluation, TARL attains the highest lemmatized terminology success rate of any submitted system (95.7%, averaged over English→Polish and Spanish→Basque), while remaining competitive on translation quality (COMET-22: 88.3, XCOMET-XXL: 83.6). On Track 2, TARL achieves 80.4% term success and 85.8 COMET-22 across three directions. A terminology-mode analysis confirms that the gains come from genuine use of the supplied dictionary rather than from instruction-following alone.

1 Introduction

Machine translation of specialized text often has to respect a fixed vocabulary. In domains such as medicine, law, finance, and engineering, a source term is expected to be rendered with one specific target term, and a translation that is otherwise fluent can still be unusable if it translates a term inconsistently or incorrectly. The WMT terminology translation task targets this setting: systems are given source text together with a bilingual glossary and are expected to produce translations that both read well and use the required terms (Semenov et al., 2025a).

Large language models now produce fluent translations across many language pairs, but they do not automatically follow a supplied glossary. A term may be paraphrased, inflected into a different form, or dropped, especially when the required target term does not fit naturally into the sentence the model would otherwise produce. This is the core difficulty of the task: satisfying a hard terminology constraint while keeping the translation accurate and natural.

Our submission, **TARL** (Terminology-Aware Agentic Repair Loop), addresses this difficulty entirely at inference time: we do not fine-tune any model. Instead, we treat terminology-aware translation as a short pipeline of four agents, each of which performs one narrow job: translating the sentence, making the required terms appear exactly, reviewing the result for remaining errors, and post-editing to fix those errors. The agents share a single rule throughout the pipeline: once a required target term is in place it is treated as locked, so later steps improve the translation around the terms rather than changing them. The pipeline is implemented with DSPy (Khattab et al., 2023), which lets us define each agent by its inputs, its output, and its instructions.

The task is organized into two tracks. In Track 1 the glossary is provided, so terminology is used as given. In Track 2 only parallel data is provided, so we first build a glossary by extracting bilingual term pairs from the parallel texts with an LLM, and then apply the same translation pipeline.

This paper describes TARL and its components. We first explain how terminology is obtained in each track and how relevant terms are retrieved for a given sentence, then describe each of the four agents in the pipeline, and how the agents' instructions are optimized with GEPA (Genetic-Pareto)(Agrawal et al., 2026), an automated prompt optimization framework. We develop and tune on the WMT25 terminology benchmark, whose document-level dictionary setting mirrors WMT26 Track 1, and we submit to the live WMT26 tracks (Charkiewicz et al., 2026). On the organizers' official WMT26 evaluation, TARL attains the highest lemmatized terminology success rate of any submitted system in Track 1 (95.7%), while remaining competitive on translation quality (COMET-22: 88.3).

2 Related Work

Terminology-constrained MT. Terminology-constrained machine translation has been studied through a series of WMT shared tasks (Alam et al., 2021; Semenov et al., 2023, 2025b), which have driven progress from sentence-level constraint injection toward document-level terminology consistency. Early neural approaches enforced term constraints by training the model to copy provided translations rather than altering the decoder (Dinu et al., 2019), and evaluation converged on character-level quality metrics (Popović, 2015, 2017) alongside dedicated terminology success-rate and consistency measures (Semenov and Bojar, 2022). In the industrial localization setting, Rosa (2026) propose a multi-agent orchestration framework for terminology-constrained translation, showing that structured agent coordination can reliably enforce domain glossaries at scale. A central finding of the WMT25 edition is that an accurate dictionary helps most for already-strong systems, and that document-level terminology remains substantially harder than the near-saturated sentence-level setting (Semenov et al., 2025b).

Systems for the WMT25 terminology task. The strongest WMT25 entries were LLM-based and fall into a few families. One line fine-tunes or reinforcement-tunes translation models toward term adherence, including dual-stage NMT-plus-LLM post-editing (Jaswal, 2025) and GRPO-based training against a terminology-adherence reward (Garcia Gilabert et al., 2025). A second line leaves model weights fixed and works at inference time, generating multiple candidate translations and reranking them with quality and terminology metrics via Pareto-optimal decoding (Guttmann et al., 2025). A third line arranges several LLMs into a workflow in which some models translate and others revise, selecting a final output with a terminology-aware quality metric (Grubišić and Korenčić, 2025). The document-level track winner combined modular source analysis, fuzzy glossary filtering, and context-carrying chunked translation over a single swappable backbone (Ingólfsdóttir et al., 2025); our terminology-retrieval design follows its fuzzy n-gram matching recipe. Our submission is inference-time and fine-tunes nothing, but unlike the generate-and-select multi-agent approach of Grubišić and Korenčić (2025), our agents are *role-specialized* (translate, verify, review, post-edit) and refine a single translation sequentially rather than serving as interchangeable candidate generators.

Multi-agent and iterative-refinement translation. A broader line of work improves LLM outputs at inference time through a generate–critique–revise loop. Self-Refine shows a single model can iteratively critique and improve its own output without additional training (Madaan et al., 2023), and Reflexion frames verbal self-feedback as a form of reinforcement (Shinn et al., 2023). In machine translation, multi-agent systems assign specialized roles to distinct agents: TRANSAGENTS mimics a human translation company for long-form literary translation (Wu et al., 2025), TACTIC organizes cognitively-motivated agents (Li et al., 2025), and MAATS decomposes quality assessment along MQM error dimensions including terminology (Wang et al., 2025). Our four-agent pipeline follows this lineage while keeping terminology strictly locked once inserted, so downstream agents refine the translation *around* the required terms rather than altering them.

Prompt optimization. Rather than hand-crafting prompts, a growing body of work treats the prompt as a learnable parameter. APE frames instruction generation and selection as a black-box search over LLM-proposed candidates (Zhou et al., 2023), while OPRO uses the LLM itself as an optimizer that iteratively proposes improved prompts conditioned on past prompt–score pairs (Yang et al., 2024). DSPy (Khattab et al., 2023) generalizes this by treating an LLM pipeline as a declarative program whose modules are compiled against a

downstream metric, with MIPROv2 jointly tuning instructions and demonstrations for multi-stage programs (Opsahl-Ong et al., 2024). We optimize our agents' instructions with GEPA , a reflective, evolutionary optimizer that reflects on execution traces and natural-language feedback and evolves a Pareto front of prompt candidates, a natural fit for the task's Pareto objective of translation quality against terminology success. Because GEPA optimizes natural-language instructions rather than weights, an instruction tuned with one backbone can be executed by another; we exploit this to optimize with a smaller model and deploy on a stronger one. As an in-loop quality signal we draw on reference-free LLM-as-judge evaluation (Kocmi and Federmann, 2023).

3 Approaches

3.1 Inflection-Aware Term Anchoring

Glossary-injection techniques control which terms appear, but not the *form* they take in context. This matters for morphologically rich targets: glossary entries are given in citation form, but their use in context may require case marking, agreement, or agglutinative suffixation. We therefore add **Inflection-Aware Term Anchoring (ITA)**, a single-pass prompting technique that treats each glossary entry as a *lexical* constraint rather than an exact surface-string constraint, targeting the WMT26 grammaticality axis (Xu and Carpuat, 2021). It needs no fine-tuning, constrained decoding, or post-editing: the guidance is placed in the document-level context given to the model. The required target lexeme must appear whenever its source term occurs, with the same lexical choice kept across repetitions while the surface form varies with local grammar. The prompt makes this explicit for the target language: it states that entries are citation forms, summarizes the language's case and affixation system and the contexts that trigger each form, and asks the model to inflect the head of a multi-word term while keeping agreement. ITA only affects output on morphologically rich targets. Analytic languages such as Chinese do not inflect glossary terms for case or agreement, so ITA reduces to plain exact-match term injection there. Chinese thus cannot distinguish ITA from a term-injection baseline on the grammaticality axis. We therefore treat ITA as a targeted contribution for morphologically rich targets and use the agentic pipeline (§3.2) as our primary, language-agnostic system for evaluation.

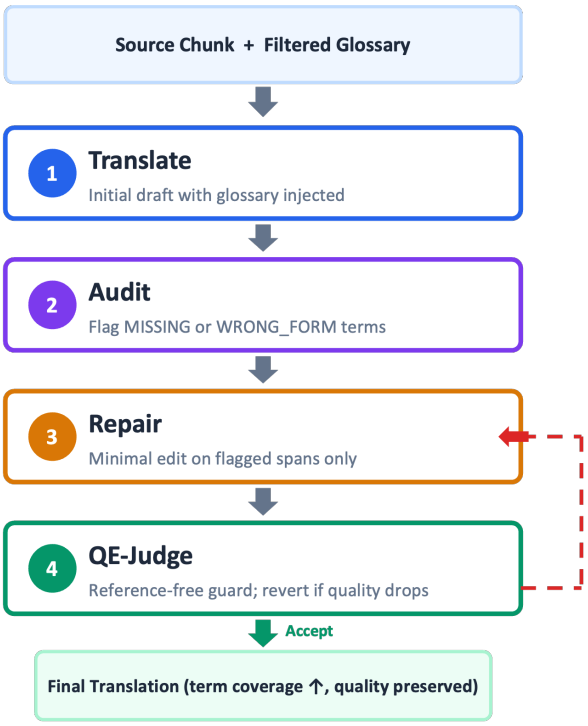


Figure 1: The four-agent terminology-aware translation pipeline. Each source chunk, together with its filtered glossary, passes through Translation, Terminology Verification (Audit), Reviewer, and Post-Editor (Repair) agents; a reference-free quality gate reverts a repair if it degrades quality. Required terms are locked once inserted, so later agents refine the surrounding text without altering them.

3.2 TARL: Agentic Terminology-Aware Translation

We approach terminology-aware translation as a pipeline of four agents that run in sequence (Figure 1). Each agent handles a specified task in the pipeline. The first agent produces a translation, the second makes sure the required terms are present, the third looks for remaining errors, and the fourth fixes those errors. The output of one agent is passed as input to the next. All four agents are implemented with the DSPy framework.Before the pipeline runs, we retrieve the terminology that applies to each source sentence and pass it to the translation agent as a list of source-to-target term pairs.

3.2.1 Terminology Retrieval

For each source sentence we look up which glossary terms it contains and build a list of (source term, target term) pairs. Retrieval is exact-match and does not use the language model. We support

two matching strategies depending on the language:

- For space-delimited languages (for example English, Spanish, Polish), we tokenize the sentence on whitespace, generate n-grams up to a fixed length, and keep any n-gram that matches a glossary key exactly. Longer matches are preferred over shorter ones, and matched spans are not allowed to overlap.

- For CJK languages (Chinese, Japanese, Korean), where text has no spaces, we match glossary keys directly as substrings of the source sentence. Longer keys are tried first so that more specific terms take priority, and matched spans are again kept non-overlapping.

When a source term maps to more than one target term, we use the first one. The result is an ordered list of term pairs formatted as "source → target", which is the form the agents receive.

3.2.2 The Four Agents

Each agent receives the source sentence and the retrieved terminology pairs, andthey run in sequence (Figure 1). The **Translation Agent**produces an initial translation that is fluent and accurate and uses each exacttarget term without paraphrasing it. The **Terminology Verification Agent**then makes minimal edits so that every required target term appears verbatim,driving the Term Success Rate toward 1.0 while preserving meaning and fluency.The **Reviewer Agent** inspects the result for remaining problems(mistranslations, omissions, grammatical errors) but treats the required terms as*locked*: it may only flag issues that do not involve them, and reports noneif every issue would require touching a locked term. Finally, the**Post-Editor Agent** applies the reviewer's fixes to produce the finaltranslation, again keeping every required term exactly as-is; terminologycompliance takes priority over other corrections.

3.3 Terminology Extraction for Track 2

Track 1 supplies a glossary as input; Track 2 gives only parallel data (source documents and their human translations). For Track 2 we therefore build the glossary ourselves before any translation takes place.We process one parallel pair at a time, prompting the LLM (Claude on Amazon Bedrock, temperature 0) with both texts and asking for the domain-specific term pairs it finds. The prompt names the domain and both languages, and requires the model to keep only domain-specific or technical terms, named entities, and multi-word terms of art (skipping generic and function words); to copy the source and target terms verbatim from their respective texts, preferring base forms and excluding surrounding punctuation or articles; and to invent nothing absent from either text. Responses are returned as a JSON array of source/target objects and parsed; pairs are processed in parallel across the corpus with a thread pool.Each pair is then validated with a loose case-insensitive substring check against the corresponding text, discarding hallucinated or reformulated terms. Surviving pairs form a glossary mapping each source term to the set of target terms observed for it.The glossary optionally passes through two filters shared with our word-alignment extraction pipeline, so both extractors produce comparable output: a *noun-phrase filter*, which POS-tags source terms with Stanza and keeps only noun phrases (a noun or proper noun with a nominal head), dropping verb-only entries; and a *deduplication filter*, which removes a shorter term subsumed by a longer term with the same target and merges source terms sharing an identical target set.The result is written in the provided-glossary format, so the rest of the pipeline treats both tracks identically from this point on.

4 Experimental Setup

Data. We report official WMT26 Terminology Shared Task (Charkiewicz et al., 2026) scores. *Track 1* supplies a bilingual dictionary and covers English→Polish (medical, mechanical-engineering) and Spanish→Basque (automotive, energy). *Track 2* provides only sample parallel bitexts and adds Traditional Chinese→English (finance). All data is document-level. We tuned on the WMT25 release, whose dictionary setting mirrors WMT26 Track 1.

Terminology modes. Track 1 is evaluated under PROPER (correct dictionary), RANDOM (decoy dictionary), and NOTERM (no dictionary); Track 2 under SAMPLE (with bitexts) and NOTERM. The PROPER−NOTERM contrast isolates the causal effect of terminology; RANDOM detects blind term insertion.

Metrics. Five metrics are reported: document-level chrF++ (Popović, 2017); lemmatized exclusive term success rate (TSR, %); and three

Track 1 (PROPER mode: en→pl, es→eu)

	chrF++			Term (lemma, %)			COMET-22			XCOMET-XXL			MetricX-24 XL (↓)		
System	Avg	en pl	es eu	Avg	en pl	es eu	Avg	en pl	es eu	Avg	en pl	es eu	Avg	en pl	es eu
gold	*100.0*	*100.0*	*100.0*	*96.4*	*93.6*	*99.1*	*96.4*	*96.3*	*96.5*	*93.2*	*96.5*	*89.8*	*2.04*	*2.26*	*1.82*
TARL	75.3	77.1	73.5	**95.7**	95.4	**96.0**	88.3	89.0	87.6	83.6	85.8	81.4	3.92	4.31	3.53
$\text{TARL}_{\text{GEPA}}$	75.8	77.6	73.9	94.1	95.1	93.1	–	–	–	–	–	–	–	–	–

Track 2 (SAMPLE mode: en→pl, es→eu, zh→en)

	chrF++				Term (lemma, %)			
System	Avg	en pl	es eu	zh en	Avg	en pl	es eu	zh en
gold	*100.0*	*100.0*	*100.0*	*100.0*	*93.4*	*94.9*	*99.1*	*86.3*
TARL	73.7	75.6	72.9	72.6	80.4	77.2	78.6	85.5
$\text{TARL}_{\text{GEPA}}$	74.8	77.0	74.8	72.7	86.7	87.3	85.8	87.2

	COMET-22				XCOMET-XXL				MetricX-24 XL (↓)			
System	Avg	en pl	es eu	zh en	Avg	en pl	es eu	zh en	Avg	en pl	es eu	zh en
TARL	85.8	88.0	84.6	84.9	81.6	85.5	76.2	83.0	4.15	4.68	4.44	3.34
$\text{TARL}_{\text{GEPA}}$	–	–	–	–	–	–	–	–	–	–	–	–

Table 1: Official WMT26 results for TARL and $\text{TARL}_{\text{GEPA}}$. **Track 1** uses the supplied dictionary; **Track 2** induces a glossary from sample bitexts. Metrics: document chrF++, lemmatized exclusive term success (%), COMET-22, XCOMET-XXL, and MetricX-24 XL (↓). *gold* = reference scored against itself. TARL achieves the highest term success of any submitted system on Track 1. Neural metrics for $\text{TARL}_{\text{GEPA}}$ are unavailable as COMET/XCOMET/MetricX are computed only through the WMT26 portal scorer API.

System	Mode	Term (%)	chrF++	COMET-22
TARL	PROPER	95.7	75.3	88.3
	RANDOM	66.8	72.4	86.5
	NOTERM	68.6	74.0	88.1
$\text{TARL}_{\text{GEPA}}$	PROPER	94.1	75.8	–
	RANDOM	68.2	73.6	–
	NOTERM	69.3	74.1	–

Table 2: **Track 1** mode analysis: PROPER (correct dictionary), RANDOM (decoy), NOTERM (no dictionary).

System	Mode	Term (%)	chrF++	COMET-22
TARL	SAMPLE	80.4	73.7	85.8
	NOTERM	68.7	73.3	86.4
$\text{TARL}_{\text{GEPA}}$	SAMPLE	86.7	74.8	–
	NOTERM	68.1	73.2	–

Table 3: **Track 2** mode analysis: with sample bitexts (SAMPLE) and without (NOTERM).

neural metrics on aligned segments—COMET-22, XCOMET-XXL, and MetricX-24 XL (↓). Systems are ranked by Pareto domination over (chrF++, TSR) per direction (Semenov et al., 2025b).

Configurations. We submit two configurations, both using Claude Opus on Amazon Bedrock with no fine-tuning: **TARL** (hand-tuned prompts) and **$\text{TARL}_{\text{GEPA}}$** (prompts optimized with GEPA (Agrawal et al., 2026) against a combined quality-plus-terminology objective). Both run on Track 1 and Track 2; for Track 2 the glossary is induced from parallel data (§3.2). In the tables, *gold* is the reference ceiling. Neural metrics for $\text{TARL}_{\text{GEPA}}$ are unavailable as they are computed exclusively through the WMT26 portal scorer API.[1]

5 Results

Table 1 reports the main metrics; Tables 2–3 the mode analyses.

Terminology adherence. On Track 1 (Table 1), TARL achieves 95.7% lemmatized term success—the highest of any submitted system, ahead of $\text{COZY}_{\text{flash}}$ (95.4%) and HW-TSC (95.3%)—consistent across morphologically rich en→pl (95.4%) and surface-transparent es→eu (96.0%). The locked-term invariant prevents downstream agents from paraphrasing required terms. On Track 2, TARL reaches 80.4%, strongest on zh→en (85.5%, near the gold ceiling of 86.3%). $\text{TARL}_{\text{GEPA}}$ improves both axes on Track 2: term

[1] $\text{TARL}_{\text{GEPA}}$ was evaluated locally with the released `wmt26-terminology` toolkit, which computes chrF++ and lemmatized term success but not the neural metrics.

success rises to 86.7% (+6.3) and chrF++ to 74.8 (+1.1), indicating optimized instructions help most on the harder glossary-induction setting. On Track 1 the trade-off reverses: chrF++ rises (75.8 vs. 75.3) but term success dips slightly (94.1% vs. 95.7%).

Quality cost. TARL's COMET-22 of 88.3 on Track 1 is within 1 point of the leaders COZY and CUNI-UFAL (89.2), and its XCOMET-XXL (83.6) and MetricX (3.92) are mid-pack. The quality cost of term-locking is modest, placing TARL on the chrF++/TSR Pareto frontier that the task ranks on.

Causal check. On Track 1 (Table 2), TARL's term success jumps from 68.6% (NOTERM) and 66.8% (RANDOM) to 95.7% (PROPER)—a 27-point gain attributable only to the supplied dictionary. The small RANDOM–NOTERM gap (1.8 pts) confirms no blind insertion. $\text{TARL}_{\text{GEPA}}$ shows the same pattern (69.3/68.2→94.1). On Track 2, sample bitexts lift TARL from 68.7% to 80.4% and $\text{TARL}_{\text{GEPA}}$ from 68.1% to 86.7%, while noterm baselines stay comparable—confirming GEPA affects glossary *use*, not baseline quality (Semenov et al., 2025b).

6 Conclusion

We presented **TARL**, our submission to the WMT26 Terminology Translation Task: a fine-tuning-free system that treats terminology-aware translation as a short pipeline of four role-specialized LLM agents (translate, verify, review, and post-edit) in which a required target term, once inserted, is locked so that later stages refine the translation around the terminology rather than altering it. Relevant terms are selected per sentence by an exact fuzzy n-gram retriever (with substring matching for CJK source), and for the sample-only track we induce a glossary from the provided parallel data before translating. The agents' instructions are optimized automatically with GEPA (Genetic-Pareto) against a combined quality-plus-terminology objective rather than being hand-written. On the official WMT26 Track 1 evaluation, TARL attains the highest lemmatized terminology success rate of any submitted system (95.7% on average), while remaining competitive on translation quality (COMET-22: 88.3). On Track 2, GEPA-optimized prompts improve both term success (86.7% vs. 80.4%) and chrF++ (74.8 vs. 73.7), demonstrating that automated prompt optimization is particularly effective when the task is hardest. A mode analysis confirms the gains come from genuine use of the supplied dictionary rather than from instruction-following alone.

References

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. 2026. Gepa: Reflective prompt evolution can outperform reinforcement learning. *Preprint*, arXiv:2507.19457.

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663.

Adrian Charkiewicz, Pinzhen Chen, Thierry Etchegoyhen, Harritxu Gete Ugarte, Kamil Guttmann, Xu Huang, David Ponce, Artur Nowakowski, Frédéric Odermatt, Arturo Oncevay, Kirill Semenov, Dawei Zhu, and Vilém Zouhar. 2026. Findings of the WMT26 terminology translation task. In *To appear in Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Javier Garcia Gilabert, Carlos Escolano, Xixian Liao, and Maite Melero. 2025. Terminology-constrained translation from monolingual data using GRPO. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1335–1343, Suzhou, China. Association for Computational Linguistics.

Ivan Grubišić and Damir Korenčić. 2025. IRB-MT at WMT25 terminology translation task: Metric-guided multi-agent approach. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1302–1334, Suzhou, China. Association for Computational Linguistics. The "MeGuMa" system.

Kamil Guttmann, Adrian Charkiewicz, Zofia Rostek, Mikołaj Pokrywka, and Artur Nowakowski. 2025. Laniqo at WMT25 terminology translation task: A multi-objective reranking strategy for terminology-aware translation via pareto-optimal decoding. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1276–1283, Suzhou, China. Association for Computational Linguistics.

Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Kári Steinn Aðalsteinsson, Róbert Fjölnir Birkisson, Sveinbjörn Þórðarson, and Þorvaldur Páll Helgason. 2025. Miðeind at WMT25 general machine translation task. In *Proceedings of the Tenth Conference on Machine Translation*, pages 577–582, Suzhou, China. Association for Computational Linguistics. The "Erlendur" terminology system.

Akshat Jaswal. 2025. It takes two: A dual stage approach for terminology-aware translation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1344–1350, Suzhou, China. Association for Computational Linguistics. The "DuTerm" system.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *Preprint*, arXiv:2310.03714.

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 (EAMT)*.

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

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *Preprint*, arXiv:2303.17651.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618.

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 (Volume 6: Industry Track)*, pages 917–926, San Diego, California, USA. Association for Computational Linguistics.

Kirill Semenov and Ondřej Bojar. 2022. Automated evaluation metric for terminology consistency in MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 450–457, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025a. Findings of the WMT25 terminology translation task: Terminology is useful especially for good MTs. In *Proceedings of the Tenth Conference on Machine Translation*, pages 554–576, Suzhou, China. Association for Computational Linguistics.

Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025b. Findings of the WMT25 terminology translation task: Terminology is useful especially for good MTs. In *Proceedings of the Tenth Conference on Machine Translation*, pages 554–576, Suzhou, China. Association for Computational Linguistics.

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

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Preprint*, arXiv:2303.11366.

George Wang, Jiaqian Hu, and Safinah Ali. 2025. Maats: A multi-agent automated translation system based on mqm evaluation. *Preprint*, arXiv:2505.14848.

Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Preprint*, arXiv:2405.11804.

Weijia Xu and Marine Carpuat. 2021. Rule-based morphological inflection improves neural terminology translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5902–5914, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. *Preprint*, arXiv:2309.03409.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *International Conference on Learning Representations (ICLR)*.

A Agent Prompts

This appendix documents the DSPy signatures (instructions and field descriptions) for each of the four agents in the agentic terminology-aware translation pipeline.

A.1 Agent 1: Translation Agent

Instruction:

> You are a professional translator. Translate the source sentence from the source language into the target language with the following strict requirements:
>
> 1. **TRANSLATION**: Produce a fluent, natural, and grammatically correct translation in the target language that preserves the full meaning of the source sentence.
> 2. **TERMINOLOGY ADHERENCE**: You MUST use the exact target terms from the provided terminology pairs wherever the corresponding source term appears in the sentence. Do NOT paraphrase, substitute, or modify these terms. This is a strict requirement.
> 3. **QUALITY**: The translation must be high-quality, idiomatic, and read naturally to a native speaker of the target language. Do not produce literal or awkward translations.
> 4. **LANGUAGE**: The entire output MUST be in the target language. Do NOT leave any words in the source language except for the required terminology terms.

Input Fields:

- `source_sentence`: The source sentence to translate.
- `source_language`: The language of the source sentence.
- `target_language`: The language to translate into.
- `terminology_pairs`: Terminology pairs (source → target). You MUST use these exact target terms in your translation when the source term appears in the sentence.

Output Field:

- `translation`: The high-quality translated sentence in the target language that strictly incorporates all applicable terminology.

A.2 Agent 2: Terminology Verification Agent

Instruction:

> You are a terminology compliance editor. Your task is to edit the given translation so that the required target terminology terms appear EXACTLY as-is in the translation.
>
> Rules:
>
> 1. For each terminology pair provided, the TARGET term MUST appear verbatim in the edited translation.
> 2. Make MINIMAL edits to the translation — only change what is necessary to include the exact target terms.
> 3. Preserve the overall meaning and fluency of the translation as much as possible.
> 4. Do NOT add, remove, or change content beyond what is needed for terminology compliance.
> 5. If a target term already appears correctly in the translation, leave it unchanged.

Input Fields:

- `translation`: The initial translation that may not contain all required terminology.
- `source_sentence`: The original source sentence for context.
- `terminology_pairs`: Required terminology pairs (source → target). Each target term MUST appear exactly as-is in the output.

Output Field:

- `edited_translation`: The edited translation where ALL required target terms appear verbatim. TSR (Term Success Rate) must be 1.0.

A.3 Agent 3: Reviewer Agent

Instruction:

> You are a professional translation reviewer. Analyze the translation for inaccuracies, errors, or quality issues.
>
> Rules:
>
> 1. Compare the translation against the source sentence for semantic accuracy.
> 2. Check for mistranslations, omissions, additions, or grammatical errors.
> 3. **CRITICAL CONSTRAINT**: You MUST NOT suggest any changes to the required terminology terms listed in the terminology pairs. These terms are LOCKED and cannot be modified, rephrased, inflected differently, or removed under any circumstance — even if they appear grammatically awkward.
> 4. Only suggest changes to parts of the translation that do NOT involve the required terminology terms. Everything else (word order, non-terminology vocabulary, grammar of surrounding words, style) is fair game.
> 5. If the only issues you find involve the required terminology, state 'No actionable issues found.' since those terms cannot be changed.
> 6. Be specific and actionable in your feedback — clearly state what can be changed and confirm that terminology must remain untouched.

Input Fields:

- `source_sentence`: The original source sentence.
- `translation`: The current translation to review.
- `terminology_pairs`: Required terminology pairs (source → target). The TARGET terms are LOCKED — you MUST NOT suggest changing, inflecting, or removing them.

Output Field:

- `review`: Review listing issues that can be fixed WITHOUT touching the required terminology terms. State 'No actionable issues found.' if all issues involve locked terminology.

A.4 Agent 4: Post-Editor Agent

Instruction:

> You are a professional translation post-editor. Fix the identified inaccuracies in the translation while STRICTLY maintaining terminology compliance.
>
> Rules:
>
> 1. Address the issues identified in the review.
> 2. NEVER remove, modify, or paraphrase any of the required target terminology terms. TSR must remain 1.0.
> 3. Improve fluency, accuracy, and naturalness where possible.
> 4. If fixing an issue would require changing a required terminology term, do NOT fix that issue — terminology takes absolute priority.
> 5. The output must be the final, polished translation.

Input Fields:

- `translation`: The current translation with potential inaccuracies.
- `source_sentence`: The original source sentence for reference.
- `terminology_pairs`: Required terminology pairs (source → target). These MUST appear exactly as-is in the output — never modify them.
- `review`: Review feedback identifying inaccuracies to fix.

Output Field:

- `final_translation`: The corrected translation that fixes inaccuracies while strictly preserving all required terminology terms.