Multi-Agent Orchestration for Terminology-Constrained Machine Translation in Industrial Localization

Emanuele Di Rosa
DATAmundi.ai
emanuele.dirosa@datamundi.ai

Abstract

Accurate terminology is a non-negotiable requirement in industrial localization processes: a single mistranslated domain term can violate contractual obligations and erode client trust. We present **AIDA$_{\textbf{term}}$**, a deployed multi-agent LLM pipeline that orchestrates four specialized agents—Analysis, Translation, Post-editing, and Review—for terminology-constrained machine translation. The system introduces terminology-aware pre-analysis, explicit glossary injection at every pipeline stage, and a reasoning-enabled Review agent. We evaluate six configurations on the WMT25 Terminology Translation benchmark (Track 1: en→de/es/ru, IT domain), enabling systematic ablation of each design choice. Our best configuration achieves 99.4% average terminology accuracy while attaining the highest ChrF2++ scores across all three language pairs, outperforming all 20 systems submitted to the shared task. Unlike other multi-agent approaches in WMT25 that rely on generate-and-select strategies, AIDA$_{\text{term}}$ is the first to apply a role-specialized sequential pipeline to terminology-constrained MT, and is deployed with native XLIFF integration for seamless CAT tool interoperability. The system processes thousands of terminology-constrained requests daily at a large localization provider.

Industry Track category: *Deployed*.

1 Introduction

Machine translation (MT) quality has improved dramatically with the advent of large language models (Jiao et al., 2023; Hendy et al., 2023; Kocmi et al., 2025b). Yet in industrial localization—where translation is delivered as a service to enterprise clients—raw translation quality is necessary but insufficient. Clients in technical domains (IT, legal, finance, life sciences) use *termbases*: curated glossaries of domain-specific terms with approved translations that must be used *verbatim* and in the right context. A product name rendered differently across deliverables, or a regulatory term translated inconsistently, carries real contractual and reputational risk.

Enforcing terminology in neural MT has been studied via constrained decoding (Dinu et al., 2019), inline annotation (Exel et al., 2020), and LLM prompting (Moslem et al., 2023). The WMT Terminology shared tasks (Alam et al., 2021; Semenov et al., 2023, 2025) have established standardized benchmarks for this problem. In the latest edition (WMT25), 13 teams submitted 20 systems across sentence-level IT translation (Track 1: en→de/es/ru). These systems employ diverse strategies including constrained prompting, fine-tuning, dual-stage NMT+LLM post-editing (Jaswal, 2025), Pareto-optimal reranking (Zhu et al., 2025), and—in the case of MeGuMa (Grubišić and Korenčić, 2025)—metric-guided multi-agent generation and selection.

Meanwhile, multi-agent LLM systems have shown promise in general translation by decomposing the task into sub-steps handled by specialized agents (Wu et al., 2024; Zhang et al., 2024a; Li et al., 2025; Briva-Iglesias et al., 2024). MeGuMa applies a multi-agent strategy to WMT25 terminology translation, but its agents serve as parallel *generators* whose outputs are ranked by composite metrics—a fundamentally different paradigm from a *sequential pipeline* where role-specialized agents progressively refine the translation through analysis, drafting, editing, and review.

This paper presents **AIDA$_{\textbf{term}}$**, a production-ready multi-agent translation system for terminology-constrained MT. AIDA$_{\text{term}}$ is deployed in production at an industrial localization provider, where it serves as the primary MT engine for clients requiring strict terminology compliance. We make the following contributions:

- We present a deployed, role-specialized multi-agent pipeline for terminology-constrained MT

that outperforms all 20 systems submitted to the WMT25 Terminology benchmark—including other multi-agent approaches—achieving the highest combined ChrF2++×Accuracy across all three language pairs (§5).

- We conduct a systematic ablation over six configurations that isolates the contribution of the Analysis agent, prompt-level terminology directives, model generation, and review with high reasoning effort enabled (§6).
- We report on production deployment at an industrial localization provider, discussing design trade-offs between quality, latency, and cost that shaped the system (§7).

Below, $\text{AIDA}_{\text{term}}$ denotes the configuration of our general-purpose AIDA Agents platform specifically tailored to terminology-constrained tasks.

To our knowledge, this is the first "architecture over models" demonstration for terminology-constrained MT: even with previous-generation models and no reasoning (V4), $\text{AIDA}_{\text{term}}$ outperforms all 20 WMT25 systems and beats the best single-agent reasoning model (o3-term-guide) while providing a cost-deterministic alternative—a fixed number of API calls, as opposed to the variable, hard-to-predict cost of reasoning-based single-agent systems.

2 Related Work

Terminology in MT. The WMT terminology tasks (Alam et al., 2021; Semenov et al., 2023, 2025) have tracked progress from constrained decoding (Dinu et al., 2019) through LLM-based approaches. In WMT25, top systems include dual-stage pipelines combining NMT with LLM post-editing (Jaswal, 2025), Pareto-optimal reranking (Zhu et al., 2025), metric-guided multi-agent selection (Grubišić and Korenčić, 2025), GRPO-aligned open LLMs (Garcia Gilabert et al., 2025), and large-scale models with difficulty-filtered training (Kocmi et al., 2025a). A key finding is that "terminology is useful especially for good MTs" (Semenov et al., 2025): high-quality base translation amplifies terminology gains.

Multi-Agent Translation. TransAgents (Wu et al., 2024) introduced role-based agents (senior editor, translator, proofreader) for literary translation. TACTIC (Zhang et al., 2024a) adds cognitive-theoretic collaboration, MAATS (Li et al., 2025) integrates MQM-based quality estimation, and GRAFT (Zhang et al., 2024b) targets document-level coherence. Zhao et al. (2025) survey the emerging landscape.

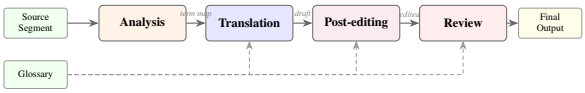

Figure 1: $\text{AIDA}_{\text{term}}$ pipeline. Solid arrows: processing flow; dashed arrows: glossary injection.

Among WMT25 terminology participants, MeGuMa (Grubišić and Korenčić, 2025) is the closest multi-agent system to ours. MeGuMa deploys open-weight models from three families (Gemma 3, Qwen 3, EuroLLM) as parallel generators, each producing candidate translations that are then revised by thinking-enabled models; the final output is selected via a composite metric combining MetricX and terminology accuracy. This *generate-and-select* paradigm differs from $\text{AIDA}_{\text{term}}$ in three key respects: (i) $\text{AIDA}_{\text{term}}$ have *distinct roles* (analysis, translation, post-editing, review) rather than serving as interchangeable generators; (ii) $\text{AIDA}_{\text{term}}$ uses frontier proprietary models in a heterogeneous cross-provider configuration (GPT-5/5.2 + Claude 4.5 Sonnet), whereas MeGuMa uses open-weight models (7–27B); and (iii) $\text{AIDA}_{\text{term}}$ enforces terminology through pipeline-integrated compliance checks at every stage, while MeGuMa enforces it through post-hoc selection among candidates. Other WMT25 systems also combine multiple components—Erlendur uses modular preprocessing, translation, and post-processing with a single backbone model, and DuTerm (Jaswal, 2025) chains an NMT model with LLM post-editing—but neither constitutes a multi-agent architecture with specialized, interacting agents.

3 System Architecture

The $\text{AIDA}_{\text{term}}$ pipeline comprises four sequential LLM-powered agents. Figure 1 illustrates the architecture and information flow.

Analysis Agent. Given a sample of the source content, the Agent identifies key textual and linguistic features of the processed material. The Agent provides guidance to the downstream operations in terms of document domain, purpose, estimated audience and author competence. This explicitly provided information helps with correct resolution of homonymy as well as actual meaning of phrases/terms in the context of the processed content by the following Agents. Concretely, the pre-analysis covers domain identification, audience

profiling and extraction of salient linguistic features (register, style, tagged placeholders), which are then passed as context to the Translation, Post-editing and Review agents.

Terminology injection. Terminology entries are retrieved from a termbase attached in the corresponding CAT project. For this task terminology has been provided at segment level within the batches (all $\text{AIDA}_{\text{term}}$ versions) as well as at a batch level (V3 version). Both variants used filtered terminology entries which are applicable to the currently processed batch. This allows us to evaluate the impact of the implicit terminological decision process by the Agents (disambiguation). In production, termbases provided by clients can be noisy (duplicates, conflicting entries, stale or out-of-domain entries), so $\text{AIDA}_{\text{term}}$ additionally performs an implicit glossary cleaning step before injection: a prompt-based relevance filter, driven by the Analysis agent, retains only the entries whose source form is attested in the current batch and flags duplicates and conflicts.

Translation Agent. Receives the source text and the Analysis agent's terminology map, and produces an initial translation. The prompt includes explicit directives to use the specified terms verbatim. We use a model from a different provider family (Claude 4.5 Sonnet) than the other agents (GPT-5/5.2), introducing beneficial diversity: systematic biases of one model family are more likely to be caught by another.

Post-editing Agent. Performs targeted corrections on the initial translation, with particular attention to terminology compliance. The prompt instructs minimal, conservative edits—this agent must not rewrite the translation but only fix errors, with each flagged term verified against the glossary.

Review Agent. Performs a dual quality gate: (1) general adequacy and fluency, and (2) terminology compliance. Any detected non-compliance is corrected. In $\text{AIDA}_{\text{term}}$ V5, the Review agent uses the GPT-5.2 model with reasoning effort enabled at the highest setting, allowing extended chain-of-thought verification of each terminology constraint before issuing its verdict.

3.1 System Configurations

We evaluate six configurations (Table 1) designed to isolate the contribution of each design decision: the Analysis agent, prompt-level terminology directives, batch vs. per-segment term processing, model generation, and reasoning-enabled review.

ID	Pipeline	Key Variation
V1	A-T-P-R	Baseline pipeline (GPT-5 / Claude 4.5S / GPT-5 / GPT-5.2)
V1.1	A-T-P-R	V1 + **stern term directive** in all prompts
V2	T-P-R	**No Analysis agent**; term. passed directly
V3	A-T-P-R	V1 + **batch term instructions** (all terms at once)
V4	A-T-P-R	**Older models**: GPT-4o / Claude 3.7S / GPT-4o / GPT-4o
V5	A-T-P-R	V1.1 + **Review reasoning** at xhigh

Table 1: System configurations. A=Analysis, T=Translation, P=Post-editing, R=Review. "S" abbreviates "Sonnet".

4 Experimental Setup

Benchmark. We evaluate on WMT25 Terminology Translation Task – Track 1 (Semenov et al., 2025): sentence-level IT-domain translation for en→de, en→ru, and en→es. Each source segment is paired with a glossary specifying approved target-language translations for domain-specific terms.

Terminology modes. Following the shared task protocol, we evaluate under three conditions: PROPER (correct glossary), RANDOM (randomly selected glossary entries), and NOTERM (no glossary).This factorial design enables causal analysis of whether a system genuinely leverages terminology or merely benefits from instruction-following.

Metrics. We adopt the official evaluation suite: ChrF2++ for overall translation quality (Popović, 2015), Accuracy (% of glossary terms correctly rendered), and Consistency (same source term always translated identically) (Semenov and Bojar, 2022). The official ranking metric is ChrF2++ $\times$ Acc (Proper mode).

Baselines. We compare against all 20 systems submitted to WMT25 Track 1. Our evaluation is post-hoc: we process the identical test data with the official evaluation scripts released by the organizers, ensuring direct comparability. Note that the models used in V4 (GPT-4o, released May 2024; Claude 3.7 Sonnet, released February 2025) predate the WMT25 test-set release (June 2025), rul-

ing out any form of implicit test-set tuning for that configuration. In addition to o3-term-guide (the top-ranked WMT25 single-agent system), we also run GPT-5 as a single agent with the same prompt templates and batch-processing setup as our V3 pipeline stage (same model, same default reasoning effort), to isolate the contribution of multi-agent orchestration from that of the underlying model; results are reported in Table 6.

Reasoning effort. Reasoning effort is configured as follows: in V5 the Review agent (GPT-5.2) runs at the highest available reasoning-effort setting (extra-high); in V1, V1.1, V2, V3 and V4, all agents use the provider default for the corresponding model. No model is fine-tuned.

5 Results

Table 2 presents results for all 27 systems on WMT25 Track 1 (Proper mode), ranked by the official ChrF×Acc metric. Results in all tables follow the convention: **bold** = column best, ***bold italic*** = column best among AIDA competitors, *italic* = column second-best. Five of six $\text{AIDA}_{\text{term}}$ configurations occupy ranks 1–4 and 6, outperforming all 20 WMT25 submissions. $\text{AIDA}_{\text{term}}$ V5 achieves ChrF×Acc of 74.5, a +4.2 point margin over the top WMT25 system (o3-term-guide, 70.3). Full per-language results across all three terminology modes are provided in Table 8 (Appendix C).

$\text{AIDA}_{\text{term}}$ configurations achieve the highest ChrF2++ and terminology accuracy simultaneously, challenging the assumed quality–accuracy trade-off and supporting Semenov et al.'s finding that terminology is most useful for already-good MTs. Figure 2 visualizes this: $\text{AIDA}_{\text{term}}$ configurations dominate the Pareto frontier.

6 Analysis

The six configurations enable three targeted ablations.

6.1 Stern Directives and Reasoning Review

The progression V1 → V1.1 → V5 isolates the effects of prompt-level terminology directives and reasoning-enabled review, holding the pipeline architecture constant.

Table 3 shows that stern directives (V1→V1.1) improve accuracy by +3.6 pp without degrading ChrF2++ or BLEU, and reasoning review (V1.1→V5) adds +1.9 pp, reaching 99.4%. Neither intervention harms translation quality.

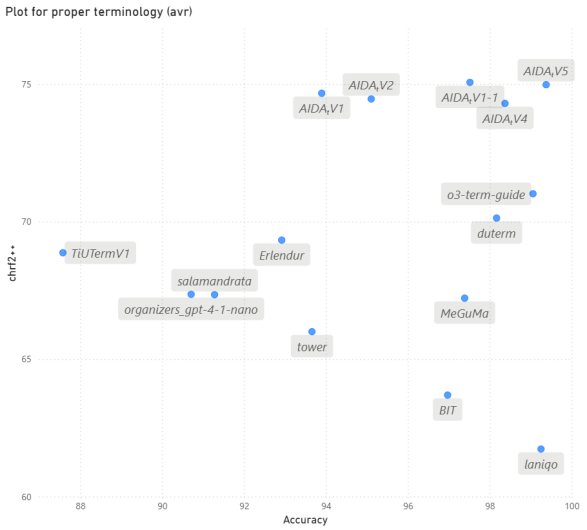

Figure 2: ChrF2++ vs. terminology accuracy (Proper mode, averaged over three language pairs). $\text{AIDA}_{\text{term}}$ configurations dominate the top-right quadrant.

6.2 Analysis Agent and Batch Processing

Table 4 reveals that removing the Analysis agent (V2) has a modest effect—the remaining agents handle terminology adequately when it is provided directly. The Analysis agent contributes primarily to translation quality (ChrF) via contextual guidance. In contrast, batch terminology processing (V3) is highly detrimental in this benchmark: segment-level measured accuracy drops by 22.4 pp. This is likely the result of the LLM agents implicitly curating the terminology list at the batch level (e.g., by removing duplicate or conflicting per-segment recommendations), so that once a term is resolved for one segment it may be silently dropped for the others where it also applies. This is a critical deployment insight: terminology must be processed per segment.

6.3 Model Generation

Comparing V1 (GPT-5/Claude 4.5 Sonnet) with V4 (GPT-4o/Claude 3.7 Sonnet) isolates the effect of model capability while holding architecture constant (Table 5). Surprisingly, V4 with older models achieves higher terminology accuracy (+4.5 pp) than V1, suggesting that instruction-following reliability may matter more than raw capability for terminology compliance. However, with stern directives and reasoning review (V5), newer models close this gap and surpass V4 on all metrics.

Architecture over models. These ablations demonstrate that $\text{AIDA}_{\text{term}}$'s gains are not attributable to using more capable LLMs. V4

#	System	en→de		en→es		en→ru		Avg
		ChrF	Acc	ChrF	Acc	ChrF	Acc	C×A
1	**AIDA$_t$ V5**	**75.1**	*99.1*	80.3	**99.4**	*69.5*	**99.6**	**74.5**
2	**AIDA$_t$ V1.1**	*74.6*	98.0	*80.8*	98.7	**69.9**	95.9	*73.3*
3	**AIDA$_t$ V4**	73.8	98.3	79.2	98.0	**69.9**	98.8	73.1
4	**AIDA$_t$ V2**	*74.6*	96.1	80.6	96.6	68.2	92.5	70.9
5	o3-term-guide	***71.6***	*99.1*	75.9	***99.1***	***65.6***	*99.0*	***70.4***
6	**AIDA$_t$ V1**	**75.1**	95.8	**80.9**	95.2	68.0	90.8	70.2
7	duterm	70.7	98.2	76.1	98.7	63.6	97.6	68.9
8	MeGuMa	67.7	96.3	72.0	97.0	61.9	98.8	65.4
9	Erlendur	69.9	93.2	74.8	94.4	63.3	91.2	64.5
10	tower	65.9	94.8	74.0	95.0	58.1	91.2	61.9
11	BIT	62.4	98.0	69.8	96.3	58.9	96.7	61.8
12	salamandrata	69.6	91.7	72.0	92.7	60.4	89.4	61.5
13	laniqo	59.8	**99.4**	68.5	98.7	56.9	***99.6***	61.2
14	org. gpt-4-1-nano	67.4	89.0	72.4	95.2	62.3	88.0	61.2
15	TiUTermV1	65.7	87.3	***77.1***	89.4	63.8	86.1	60.4
16	CommandA_MT	67.6	86.9	70.7	81.9	59.3	70.7	52.9
17	**AIDA$_t$ V3**	70.8	73.1	76.5	70.4	66.8	71.1	51.0
18	TiUTermV0	61.0	71.1	69.0	75.2	58.3	76.8	46.7
19	LC-primary	61.2	70.7	68.9	74.1	54.2	65.8	43.3
20	LC-2	61.0	70.7	67.7	73.6	53.7	65.6	42.7
21	LC-3	61.0	70.7	67.7	73.6	53.7	65.6	42.7
22	CurTermNLLB	60.3	79.0	69.1	76.5	51.0	34.6	39.4
23	ContexTerm	40.2	79.9	53.7	68.5	51.5	67.6	34.6
	Systran_gen_ft†	–	–	71.1	44.1	–	–	31.4
	EuroLLM-ft†	–	–	63.5	38.9	–	–	24.7
	MarianMT-ft†	–	–	65.6	17.5	–	–	11.5
	TranssionMT	1.6	0.0	1.3	0.0	47.8	33.2	5.3
	Δ (V5 − best WMT25)	+3.5	-0.3	+3.8	+0.3	+4.3	+0.0	+4.2

Table 2: Complete WMT25 Terminology Track 1 results (Proper mode), ranked by the official ChrF×Acc/100 metric averaged across language pairs. Highlighting convention as in §5. † = es-only. The Δ row shows the gain of our best system (V5) over the best WMT25 competitor (o3-term-guide).

	ChrF	BLEU	Acc	Cons
V1 (baseline)	74.7	*52.8*	93.9	*87.3*
V1.1 (+stern)	**75.1**	*52.8*	*97.5*	**88.1**
V5 (+reason.)	*75.0*	**52.9**	**99.4**	87.1
Δ V1.1−V1	+0.4	0.0	+3.6	+0.8
Δ V5−V1.1	−0.1	+0.1	+1.9	−1.0

Table 3: Progression from V1 to V5 (avg. across language pairs, Proper mode). Stern directives yield +3.6 pp accuracy; reasoning review adds +1.9 pp, reaching 99.4%.

	ChrF	BLEU	Acc	Cons
V1 (with Anal.)	**74.7**	**52.8**	*93.9*	*87.3*
V2 (no Anal.)	*74.5*	*52.7*	**95.1**	86.7
V3 (batch terms)	71.4	50.9	71.5	**89.8**
Δ V1−V2	+0.2	+0.1	−1.2	+0.6
Δ V1−V3	+3.3	+1.9	+22.4	−2.5

Table 4: Ablation of the Analysis agent (avg., Proper). V2 removes the agent; V3 sends all terms in a batch.

	ChrF	BLEU	Acc	Cons
V1 (current gen.)	**74.7**	**52.8**	*93.9*	**87.3**
V4 (prev. gen.)	*74.3*	*52.5*	**98.4**	**87.3**
Δ V1−V4	+0.4	+0.3	−4.5	0.0

Table 5: Model generation comparison (avg., Proper).

with previous-generation models (ChrF×Acc = 73.1) already outperforms the top WMT25 system (o3-term-guide, 70.4), showing that the pipeline architecture—terminology pre-analysis, stage-wise glossary injection, and review-reject feedback—is the primary driver. Prompt-level directives (+3.1 pp) and reasoning-enabled review (+1.2 pp) yield cumulative gains without changing models, confirming that orchestration decisions matter more than model scale. Noteworthy for deployed systems, V4—a four-stage pipeline without reasoning—already performs better than a single-agent reasoning model (o3-term-guide), while offering a cost-deterministic alternative: the number of API calls is fixed, whereas reasoning models incur variable costs that depend on processing time and are difficult to predict a priori. Finally, the

Comparison	BLEU	ChrF	Acc	Cons	C×A
V3 − GPT-5 (single)	+5.25	+2.87	−1.59	+0.74	+0.96
V5 − o3-term-guide	+6.46	+3.97	+0.32	−0.65	+4.17

Table 6: Deltas of $\text{AIDA}_{\text{term}}$ (Proper mode) over strong single-agent baselines (Proper mode, avg. over en→de/es/ru). V3 uses GPT-5 with default reasoning, matching the single-agent GPT-5 run.

consistently strong results obtained with different combinations of proprietary models suggest that open-weight models used within the same pipeline should achieve comparable performance, which we leave for future work.

6.4 Single-Agent Baselines

To directly quantify the contribution of multi-agent orchestration, we compare each of our $\text{AIDA}_{\text{term}}$ configurations against a strong single-agent reference: o3-term-guide (the top-ranked single-agent WMT25 submission) and GPT-5 used as a single agent with the same prompt templates and default reasoning-effort as our V3 pipeline stage. Table 6 reports the deltas averaged across the three language pairs.

V5 scores higher than both single-agent baselines on the official ranking metric (ChrF×Acc), with a +4.17 point margin over o3. The V3 vs. GPT-5 single-agent contrast—same base model, same prompts, same reasoning-effort—directly demonstrates that multi-agent orchestration, not the underlying model capability, drives the quality gains.

6.5 Terminology Mode Analysis

The gap between Proper and NoTerm modes for $\text{AIDA}_{\text{term}}$ V5 (+6.4 ChrF2++ avg.; see Table 8 in Appendix) confirms that the system genuinely leverages terminology to improve translation quality. Under the Random condition, accuracy against the correct terms drops to 50.1, which seems to suggest that introducing terminological "noise" by suggesting irrelevant terminology moves the focus away from the critical in-domain terms. This is mitigated in production to a large extent by proper terminology management processes.

7 Industrial Deployment

$\text{AIDA}_{\text{term}}$ is deployed in production at a localization service provider, processing terminology-constrained translation in IT, legal, and financial domains. Our multi-agent approach is applied also beyond terminology constrained tasks, and our internal experimental analysis on 11 languages in WMT24++ benchmark, shows comparable BLEU score results to the top scored system, testifying its general-purpose applicability. We summarize key design decisions and lessons learned.

Latency management. The four-agent pipeline introduces sequential latency (0.25–0.75s per segment vs. 0.25–0.5s for single-model MT). In production: (i) segments are processed in parallel in batches (of token count dependent size) across a request; (ii) Analysis agent outputs are cached and reused by all the agents along with preselected terminology; (iii) V2 (no Analysis) serves as a low-latency fallback for non-critical content. Please note that our approach allows to set the batch size, i.e. the number of segments in a translation batch, thus allowing to leverage the model context window and generalize, in some cases, up to document-level translation.

Cost structure. The pipeline requires $\sim 4\times$ the token throughput of single-model MT, but the resulting 40–60% reduction in human post-editing time for terminology errors more than offsets the increased API cost.

Human evaluation. Beyond the terminology-focused WMT25 evaluation, we report preliminary results validating the general translation quality of the underlying AIDA Agents platform on an industrial benchmark comprising 96 segments per language pair across six language pairs (English to Japanese, Chinese, French, Korean, German, Spanish), drawn from software documentation, user manuals, and technical specifications. Three professional linguists per language blindly evaluated translations on a 5-point scale (1=Poor to 5=Excellent). Table 7 compares AIDA Agents against Microsoft Translator (raw NMT output), showing the percentage of segments rated Good or Excellent ($\geq$4). The key insight is that 70–98% of segments are rated publication-ready without any human post-editing. On another, more extensive, human evaluation, comprising 4,418 segments from a difficult terminology-constrained, hyper-specialized scientific and engineering domain from English to Polish, every segment was processed through the standard industry two-stage human quality gate used in our production workflow: (i) a first professional translator performs post-editing on the MT output, and (ii) a second, independent professional

Lang.	MS NMT	AIDA	Δ
German	82.7%	**98.0%**	+15.3%
Spanish	92.9%	**94.9%**	+2.0%
French	60.2%	**90.8%**	+30.6%
Korean	73.4%	**81.6%**	+8.2%
Chinese	62.2%	**77.6%**	+15.4%
Japanese	12.2%	**69.4%**	+57.2%

Table 7: Human evaluation: % segments rated Good/Excellent. 70–98% publication-ready without human post-editing.

linguist performs a blind quality review of the post-edited text. No sampling was involved—all 4,418 segments passed through both stages. 42.4% of segments were kept unchanged by the first translator (i.e. the MT output was deemed publication-ready), and 34.9% were kept unchanged by the second reviewer (i.e. the post-edited translation required no further change). On the segments that were edited, ChrF2++ was 82.3 after the first stage and 78.4 after the quality-review stage, indicating that the human corrections were minor surface-level edits rather than substantive rewrites. Note that while the system has been extensively validated across numerous client projects, contractual restrictions prevent public release of broader evaluation data. Results mentioned here provide a reference on the performance of our methodology in real-world industrial applications.

Configuration selection. Based on our ablation results, for best quality we recommend V5 for high-value content requiring maximum terminology compliance, as a cost-effective alternative V1.1 is recommended for standard production (best quality–latency trade-off), and V2 for high-throughput scenarios.

XLIFF and CAT tool integration. AIDA Agents operates natively on XLIFF (XML Localization Interchange File Format), the industry-standard interchange format. Segments are extracted from XLIFF with terminology and additional references, enabling drop-in integration with existing localization workflows. The production system additionally supports RAG-based injection of translation memories and style guides at the Analysis and Translation stages, enabling domain adaptation without fine-tuning.

Research access. To facilitate reproducibility and further research on multi-agent terminology-constrained translation, access to the $\text{AIDA}_{\text{term}}$ system will be made available for research purposes. We also release prompt templates for all four agents and all six configurations, together with the raw and post-processed output translations used to compute our results, at https://github.com/emanueledirosa/aida_t-acl2026-industrytrack.

8 Conclusion

We presented $\text{AIDA}_{\text{term}}$, a deployed multi-agent pipeline for terminology-constrained MT with native XLIFF/CAT tool integration. Five of six configurations outperform all 20 WMT25 submissions—including other multi-agent approaches—on the official ranking metric, with V5 achieving 99.4% terminology accuracy alongside the highest ChrF2++. Our ablation demonstrates that the pipeline architecture—not model choice—drives performance: even with previous-generation models and no reasoning (V4), $\text{AIDA}_{\text{term}}$ surpasses all WMT25 systems, and in particular beats the best single-agent reasoning model (o3-term-guide) while providing a cost-deterministic alternative with a fixed number of API calls. Orchestration decisions yield cumulative gains without changing models. The system is deployed in industrial localization.

Limitations

Our evaluation is made on WMT25 Track 1. Results are expected to generalize to document-level translation (Track 2), since it is possible to set the system batch size, but its experimental evaluation is not presented here. We also leave a systematic analysis of document-level terminological consistency for future work; in production we observe that both document-level terminology consistency and translation fluency improve with larger, coherent batches, possible in modern context windows. The applicability of the system on other domains, or lower-resource languages, is expected, and our preliminary results are promising, but such results are not fully presented here. The system results presented rely on proprietary (publicly available) LLMs, determining a reproducibility cost, but its architecture is general-purpose and allows open weight models to be used instead. API costs scale linearly with the number of agents. We evaluate post-hoc on released test data rather than as official task participants, though we use identical data and evaluation scripts and our results are reproducible.

Ethics Statement

This work uses publicly available WMT25 benchmark data. Proprietary models are accessed via commercial APIs under standard terms. Production data in §7 is anonymized and aggregated; no individual translator data is reported. We acknowledge that automated MT pipelines affect translator workflows. We position the system as an assistive tool: human translators review and approve all outputs in production, and the system's terminology compliance reduces their corrective burden rather than replacing their expertise.

Acknowledgments

We acknowledge and thank Piotr Peszynski for his outstanding contribution to the experimental evaluation phase and the implementation effort behind $\text{AIDA}_{\text{term}}$. We also thank DATAmundi.ai for supporting this research, and the anonymous ACL reviewers for their constructive feedback.

References

Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Laurent Besacier, and 1 others. 2021. Findings of the WMT 2021 shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation (WMT)*. Association for Computational Linguistics.

V. Briva-Iglesias, C. Campolargo, S. Coyne, and A. Way. 2024. Legal translation: The potential of multi-agent LLMs for improved efficiency and quality. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, pages 421–430. European Association for Machine Translation.

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. Association for Computational Linguistics.

Miriam Exel, Bianka Buschbeck, Laurie Brandt, and Simona Doneva. 2020. Terminology-constrained neural machine translation at SAP. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 271–280.

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 (WMT)*, 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 (WMT)*, pages 1302–1334, Suzhou, China. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, and 1 others. 2023. How good are GPT models at machine translation? A comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Akshat Singh Jaswal. 2025. It takes two: A dual stage approach for terminology-aware translation. In *Proceedings of the Tenth Conference on Machine Translation (WMT)*, Suzhou, China. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Tom Kocmi, Arkady Arkhangorodsky, Alexandre Berard, and 1 others. 2025a. Command-A-translate: Raising the bar of machine translation with difficulty filtering. In *Proceedings of the Tenth Conference on Machine Translation (WMT)*, pages 789–799, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, and 1 others. 2025b. Findings of the WMT25 general machine translation shared task. In *Proceedings of the Tenth Conference on Machine Translation (WMT)*. Association for Computational Linguistics.

H. Li, J. Chen, and Q. Liu. 2025. MAATS: A multi-agent automated translation system based on MQM evaluation. *arXiv preprint arXiv:2505.14848*.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237.

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. 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. Association for Computational Linguistics.

Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025. Findings of the WMT25 terminology translation task: Terminology is useful especially for good MTs. In *Proceedings of the Tenth Conference on Machine Translation (WMT)*, 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 (WMT)*, pages 663–671, Singapore. Association for Computational Linguistics.

Minghao Wu, Yuan He, Yufei Zhang, Xiaojun Xu, Maosong Sun, Longyue Huang, and Lemao Liu. 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 245–256, Miami, Florida, USA. Association for Computational Linguistics.

Jiawei Zhang, Yuzhi Tian, Junjie Yin, Min Yang, and Bing Qin. 2024a. TACTIC: Translation agents with cognitive-theoretic interactive collaboration. *arXiv preprint arXiv:2506.08403*.

Liang Zhang, Siyuan Peng, and Hongbin Liu. 2024b. GRAFT: A graph-based flow-aware agentic framework for document-level machine translation. *arXiv preprint arXiv:2507.03311*.

Yang Zhao, Jiajun Liu, and Hao Wang. 2025. Are AI agents the new machine translation frontier? *arXiv preprint arXiv:2504.12891*.

Lichao Zhu, Maria Zimina-Poirot, Stephane Patin, and Cristian Valdez. 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 (WMT)*, Suzhou, China. Association for Computational Linguistics.

A Per-Language Scatter Plots

Figures 3 and 4 show per-language ChrF2++ vs. terminology accuracy scatter plots for the Proper terminology condition. $\text{AIDA}_{\text{term}}$ configurations consistently occupy the top-right region across all language pairs.

B Terminology Mode Analysis

Figure 5 shows a plot comparing the range of ChrF2++ results across terminology modes.

C Full Results Across All Systems and Modes

Table 8 presents results for all systems across all three terminology modes (Proper, Random, NoTerm), averaged over language pairs. Systems with partial language coverage are marked; per-language breakdowns for Proper mode appear in Table 9.

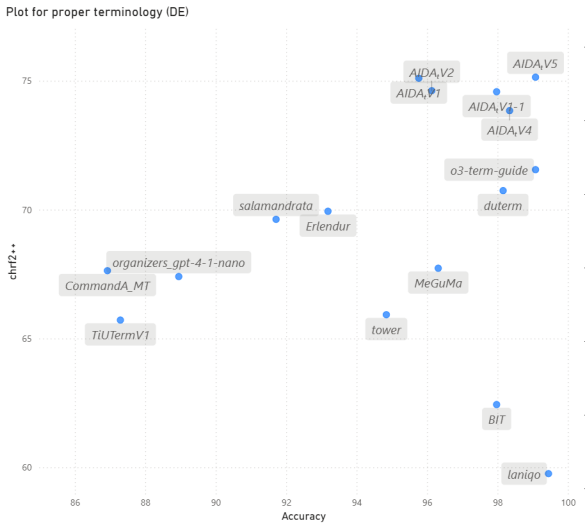

Figure 3: ChrF2++ vs. accuracy for en→de (Proper mode).

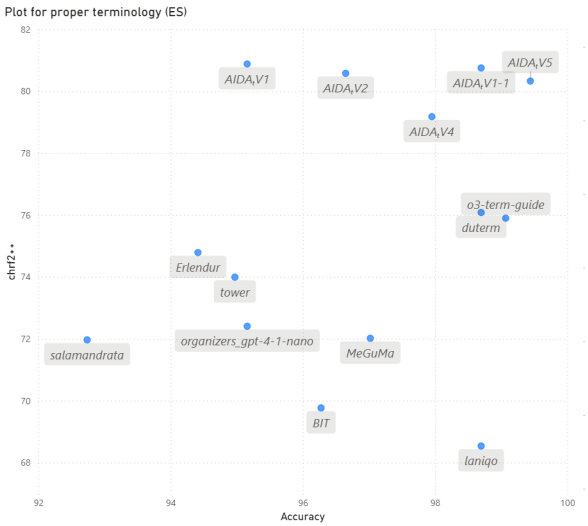

Figure 4: ChrF2++ vs. accuracy for en→es (Proper mode).

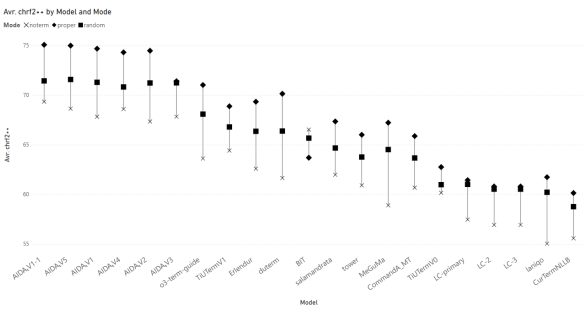

Figure 5: Average ChrF2++ across terminology modes. $\text{AIDA}_{\text{term}}$ configurations (leftmost) maintain high quality regardless of condition. Diamond: Proper; square: Random; cross: NoTerm.

		Proper			Random			NoTerm		
#	System	ChrF	Acc	Cons	ChrF	Acc	Cons	ChrF	Acc	Cons
1	**AIDA$_t$ V5**	*75.0*	**99.4**	87.1	**71.6**	50.1	87.4	*68.6*	48.5	88.3
2	**AIDA$_t$ V1.1**	**75.1**	97.5	88.1	*71.4*	49.6	87.9	**69.3**	47.1	88.7
3	**AIDA$_t$ V4**	74.3	98.4	87.2	70.8	50.9	**89.1**	*68.6*	48.6	89.0
4	**AIDA$_t$ V2**	74.5	95.1	86.7	71.2	49.9	87.4	67.3	46.8	*89.3*
5	o3-term-guide	***71.0***	99.1	87.7	68.1	49.2	***88.3***	63.6	44.4	**89.5**
6	**AIDA$_t$ V1**	74.7	93.9	87.0	71.3	48.2	87.7	67.8	45.4	88.9
7	duterm	70.1	98.2	87.3	66.4	46.6	86.6	61.6	42.9	86.9
8	MeGuMa	67.2	97.4	***88.6***	64.5	46.7	87.1	58.9	40.1	86.3
9	Erlendur	69.3	92.9	86.7	66.4	44.4	86.2	62.6	42.3	87.1
10	tower	66.0	93.7	88.4	63.8	44.3	87.4	60.9	40.9	88.3
11	BIT	63.7	97.0	87.8	65.7	**80.5**	87.9	***66.5***	**96.7**	87.9
12	salamandrata	67.3	91.3	87.4	64.7	48.2	87.4	62.0	44.4	87.9
13	laniqo	61.7	***99.3***	87.6	60.2	42.7	82.3	55.0	36.9	82.2
14	org. gpt-4-1-nano‡	67.4	90.7	87.5	–	–	–	–	–	–
15	TiUTermV1	68.9	87.6	86.7	***66.8***	*54.6*	85.1	64.4	*52.1*	85.2
16	CommandA_MT	65.9	79.9	86.6	63.7	45.8	***88.3***	60.7	43.0	87.7
17	**AIDA$_t$ V3**	71.4	71.5	**89.8**	71.2	51.7	*88.5*	67.8	45.4	88.9
18	TiUTermV0	62.7	74.4	86.4	61.0	49.6	84.9	60.2	49.1	85.2
19	LC-primary	61.4	70.2	85.4	61.0	38.6	85.4	57.5	36.5	84.7
20	LC-2	60.8	70.0	85.8	60.5	38.5	85.7	56.9	36.3	85.0
21	LC-3	60.8	70.0	86.0	60.5	38.5	84.9	56.9	36.3	85.3
22	CurTermNLLB	60.1	63.4	88.0	58.8	36.1	84.1	55.6	34.2	85.7
23	ContexTerm	48.5	72.0	81.9	48.2	24.6	80.0	45.7	22.4	79.2
	Systran_gen_ft†	71.1	44.1	88.1	71.1	44.1	88.6	71.1	44.1	88.2
	EuroLLM-ft†	63.5	38.9	82.5	63.5	38.9	83.1	63.5	38.9	82.8
	MarianMT-ft†	65.6	17.5	54.1	68.9	48.8	85.1	68.9	48.8	86.4
	TranssionMT	16.9	11.1	57.0	16.9	11.1	55.7	16.9	11.1	58.3

Table 8: All systems across three terminology modes (averaged over de/es/ru). ChrF = ChrF2++; Acc = terminology accuracy (%); Cons = consistency (%). Ranked by Proper ChrF×Acc. † = es-only. ‡ = Proper only.

		en→de			en→es			en→ru		
#	System	ChrF	Acc	Cons	ChrF	Acc	Cons	ChrF	Acc	Cons
1	**AIDA$_t$ V5**	**75.1**	*99.1*	86.8	80.3	**99.4**	84.4	*69.5*	**99.6**	90.1
2	**AIDA$_t$ V1.1**	*74.6*	98.0	87.5	*80.8*	98.7	85.6	**69.9**	95.9	*91.1*
3	**AIDA$_t$ V4**	73.8	98.3	86.0	79.2	98.0	84.8	**69.9**	98.8	90.8
4	**AIDA$_t$ V2**	*74.6*	96.1	86.8	80.6	96.6	84.7	68.2	92.5	88.5
5	o3-term-guide	***71.6***	*99.1*	86.1	75.9	***99.1***	86.7	***65.6***	*99.0*	90.4
6	**AIDA$_t$ V1**	**75.1**	95.8	86.6	**80.9**	95.2	85.6	68.0	90.8	88.8
7	duterm	70.7	98.2	86.3	76.1	98.7	86.0	63.6	97.6	89.5
8	MeGuMa	67.7	96.3	88.6	72.0	97.0	86.9	61.9	98.8	90.2
9	Erlendur	69.9	93.2	86.3	74.8	94.4	83.8	63.3	91.2	90.0
10	tower	65.9	94.8	86.8	74.0	95.0	***87.6***	58.1	91.2	***90.7***
11	BIT	62.4	98.0	86.9	69.8	96.3	86.8	58.9	96.7	89.8
12	salamandrata	69.6	91.7	86.4	72.0	92.7	87.3	60.4	89.4	88.6
13	laniqo	59.8	**99.4**	***89.3***	68.5	98.7	85.6	56.9	***99.6***	87.9
14	org. gpt-4-1-nano	67.4	89.0	86.3	72.4	95.2	86.3	62.3	88.0	90.0
15	TiUTermV1	65.7	87.3	85.9	***77.1***	89.4	85.7	63.8	86.1	88.5
16	CommandA_MT	67.6	86.9	87.5	70.7	81.9	84.5	59.3	70.7	87.8
17	**AIDA$_t$ V3**	70.8	73.1	**89.6**	76.5	70.4	**87.7**	66.8	71.1	**92.2**
18	TiUTermV0	61.0	71.1	85.6	69.0	75.2	85.0	58.3	76.8	88.6
19	LC-primary	61.2	70.7	85.8	68.9	74.1	83.6	54.2	65.8	87.0
20	LC-2	61.0	70.7	85.7	67.7	73.6	85.4	53.7	65.6	86.2
21	LC-3	61.0	70.7	85.7	67.7	73.6	85.6	53.7	65.6	86.7
22	CurTermNLLB	60.3	79.0	87.6	69.1	76.5	87.5	51.0	34.6	88.8
23	ContexTerm	40.2	79.9	85.8	53.7	68.5	75.6	51.5	67.6	84.4

Table 9: Per-language results for Proper mode (top 23 full-coverage systems). ChrF = ChrF2++; Acc = terminology accuracy (%); Cons = consistency (%). Ranked by avg. ChrF×Acc.