

AIDA Agents: A Multi-Agent Translation Platform with Context-Aware Quality Control

Emanuele Di Rosa Piotr Peszynski

DATAmundi.ai

{emanuele.dirosa,piotr.peszynski}@datamundi.ai

Abstract

We present AIDA Agents, a multi-agent translation platform that orchestrates LLM-based agents—for translation, rating, post-editing, and re-rating—delivering context-aware translations without model fine-tuning. Optional retrieval-augmented generation (RAG) injects translation memories, terminology, and style guidelines at every pipeline stage. On WMT24++ (Deutsch et al.(2025)) (11 languages), AIDA Agents outperforms all systems on 10 of 11 pairs. On an industrial benchmark, 70–98% of segments are publication-ready without human post-editing. The platform is deployed with native XLIFF integration.

1 Introduction

Enterprise localization teams face a persistent trade-off: NMT engines require expensive fine-tuning and substantial post-editing, while LLMs show quality gains but lack workflow integration. Recent research on multi-agent translation systems (Wang et al.(2025)) demonstrates the potential of agent-based pipelines but lack production integration. AIDA Agents bridges this gap by orchestrating specialized LLM agents in a pipeline with native XLIFF integration. The platform is currently deployed internally at DATAmundi to support production localization for enterprise clients, offered as a proprietary service through client engagements, with potential, public commercial availability in the upcoming months. This paper describes the system, deployment configurations, and preliminary evaluation on general-purpose, WMT24++ translation benchmark, and an industrial one.

© 2026 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

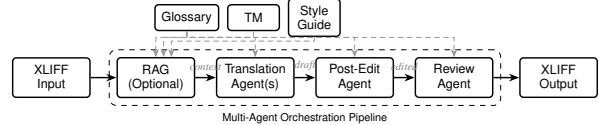


Figure 1: AIDA Agents pipeline. Dashed arrows: context injection via prompt refining and optional RAG.

2 System Description

Figure 1 illustrates the pipeline. Source content arrives from the integrated CAT tool; an optional RAG stage retrieves TM matches (via multilingual embeddings), glossary entries with approved/forbidden terms, and style instructions—enabling domain adaptation without fine-tuning. This context is injected at every pipeline stage, ensuring consistent application of client requirements. Up to three parallel translator agents produce candidates; an Arbiter/Rater, not represented in figure 1, enabled when multiple parallel translations are generated, evaluates each on MQM-inspired criteria (accuracy, fluency, terminology, style) with customizable weighted scoring, producing an explainable quality report, and selects the best one. A Post-Editor Agent refines it using the report, all candidates, and guidelines. A second Arbiter/Rater re-evaluates and produces the final output, returned to the CAT tool with XLIFF structure preserved.

Three product tiers address different quality-cost trade-offs: **AIDA LLM** (3 calls/segment): 1 translator + 1 post-editor + 1 rater; best quality-cost balance. **AIDA MultiLLM** (up to 6 calls): up to 3 parallel translators + rater + post-editor + re-rater; maximum quality through diversity. **AIDA NMT** (3 calls): 1 NMT engine + 1 LLM post-editor + 1 LLM rater; lowest cost, terminology can be injected at the post-editor step. The platform supports XLIFF format and CAT tool connectors.

System	JA	ZH	DE	FR	ES	IT	IN	NO	VI	SW	PT
AIDA MultiLLM	27.9	47.2	31.5	40.4	46.1	43.6	34.8	45.1	37.0	41.5	42.0
AIDA LLM	29.6	41.5	33.7	40.1	44.5	44.1	35.9	45.5	37.1	43.6	41.7
AIDA NMT	30.2	48.1	34.7	38.7	44.8	43.2	36.6	47.1	37.3	42.8	42.6
GPT-4o	25.7	42.8	33.8	39.2	41.3	40.5	33.2	45.1	34.3	44.9	41.6
DeepL	26.3	39.2	33.4	38.2	45.5	42.9	32.5	44.1	—	44.1	41.3
Google Tr.	25.3	41.5	34.5	37.2	43.0	41.5	33.5	43.5	34.9	41.7	42.2
Gemini-1.5-Pro	23.7	45.7	31.8	36.0	39.4	37.7	31.4	40.9	29.7	38.0	39.5
Claude-3.5	23.3	38.6	30.8	37.7	39.0	38.8	31.4	40.5	30.8	41.9	40.6
Microsoft Tr.	22.1	35.8	31.2	31.7	41.2	41.2	30.9	46.1	30.9	42.8	40.9
OpenAI o1	23.6	38.7	31.8	33.3	41.2	36.2	30.0	43.0	31.0	40.9	39.0
Unbabel Tower	23.5	40.6	32.0	37.4	41.3	41.0	—	—	—	—	41.2
Δ AIDA vs. Best	+3.9	+2.4	+0.2	+1.2	+0.6	+1.2	+3.1	+1.0	+2.4	-1.3	+0.4

Table 1: WMT24++ BLEU (EN→target). AIDA Agents best on 10/11 languages. No TM/terminology available. Results for all 16 systems available upon request.

3 Experimental Results

3.1 WMT24++ Benchmark (11 Languages)

Table 1 reports BLEU on WMT24++ (Deutsch et al.(2025))—960 segments, 4 domains, no TM/terminology—isolating the multi-agent orchestration effect. AIDA Agents achieves the top score on 10/11 languages (up to +3.9 on Japanese). With RAG-injected resources, gains are larger (§3.2).

3.2 Industrial Benchmarks

Table 2 reports results on 96 segments per language (technical IT documentation, full TM/terminology/style guides, 3 linguists per language, blind 5-point scale) previously translated in a project by a competitor company involving human professional post-editors. Our goal is to show preliminary results about the added value of our automated translation system compared to a full human-in-the-loop post-editing process.

Lang.	BLEU			Good/Exc. (%)		
	MS	Comp.	AIDA	MS	Comp.	AIDA
JA	24.0	52.8	59.3	12.2	85.7	69.4
ZH	24.9	43.1	45.2	62.2	82.7	77.6
FR	42.2	47.6	48.8	60.2	94.9	90.8
KO	48.1	54.2	56.9	73.4	82.9	81.6
DE	55.1	56.1	57.0	82.7	91.9	98.0
ES	58.5	64.4	65.3	92.9	95.9	94.9

Table 2: Industrial benchmark. MS = Microsoft NMT; Comp. = commercial system + human post-editing; AIDA = AIDA MultiLLM (automated).

AIDA MultiLLM achieves the highest BLEU on all six languages; 70–98% of segments are publication-ready without human intervention, even surpassing the human-post-edited competitor for

German (98.0% vs. 91.9%). Even though AIDA Agents workflow is automated suggesting a potential 50–70% reduction in post-editing effort, professional human post-editing is fully integrable. The experimental evaluation reported here focuses on general-purpose translation; an extended evaluation on terminology-constrained tasks and a specialized architecture have been presented in (Di Rosa(2026)).

4 Conclusion

Unlike standalone NMT engines or research multi-agent systems, AIDA Agents is a deployed orchestration platform with native CAT/TMS integration requiring no fine-tuning. It can incorporate any NMT or LLM backend and enables immediate deployment on new domains and languages. It consistently outperforms state-of-the-art MT across 11 languages on WMT24++ and delivers 70–98% publication-ready output on industrial data, even surpassing human-post-edited baselines for several language pairs. A live demo can be presented at the conference.

References

- Daniel Deutsch et al. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of ACL 2025*, pages 12257–12284.
- Emanuele Di Rosa. 2026. Multi-agent orchestration for terminology-constrained machine translation in industrial localization. In *Proc. of ACL 2026: Industry Track*. To appear.
- George Wang et al. 2025. MAATS: A multi-agent automated translation system based on MQM evaluation. arXiv:2505.14848.