MAATS: A Multi-Agent Automated Translation System Based on MQM Evaluation

Anonymous submission

Abstract

We present **MAATS**, a **M**ulti **A**gent **A**utomated **T**ranslation **S**ystem that leverages the Multidimensional Quality Metrics (MQM) framework as a fine-grained signal for error detection and refinement. MAATS employs multiple specialized AI agents, each focused on a distinct MQM category (e.g., Accuracy, Fluency, Style, Terminology), followed by a synthesis agent that integrates the annotations to iteratively refine translations. This design contrasts with conventional single-agent methods that rely on self-correction.

Evaluated across diverse language pairs and state-of-the-art Large Language Models, MAATS outperforms zero-shot and single-agent baselines with statistically significant gains in both automatic metrics and human assessments. It excels particularly in semantic accuracy, locale adaptation, and linguistically distant language pairs. Qualitative analysis highlights its strengths in multi-layered error diagnosis, omission detection across perspectives, and context-aware refinement. By aligning modular agent roles with interpretable MQM dimensions, MAATS narrows the gap between black-box LLMs and human translation workflows, shifting focus from surface fluency to deeper semantic and contextual fidelity.

Introduction

While both multi agent systems and machine translation have seen rapid progress (Manakhimova et al. 2023; Peng et al. 2023; Jiao et al. 2023), the use of explicitly specialized agents assigned to distinct evaluation tasks remains underexplored. Despite advances that enable LLMs to rival dedicated MT systems, top-performing models still produce subtle mistranslations, omissions, stylistic errors, and inconsistencies (Freitag et al. 2021a; Yan et al. 2024).

Researchers are drawing inspiration from professional human translation workflows, which typically involve a multi-stage process with collaboration between translators and editors (Yan et al. 2014; Qian and Kong 2024; Wu et al. 2024). This insight has led to refinement approaches for LLMs, including single-agent methods where a model attempts to critique and correct its own output (Feng et al. 2024). Although such self-refinement can reduce errors, a single model can struggle to identify all issues due to potential biases or a lack of diverse expertise (Xu et al. 2024; Kamoi et al. 2024), which have implications on errors in machine translation. Recent work shows that assigning complementary roles to multiple LLM agents improves performance in complex tasks such as code generation, game playing, and long-horizon planning (Manakhimova et al. 2023; Peng et al. 2023; Jiao et al. 2023).

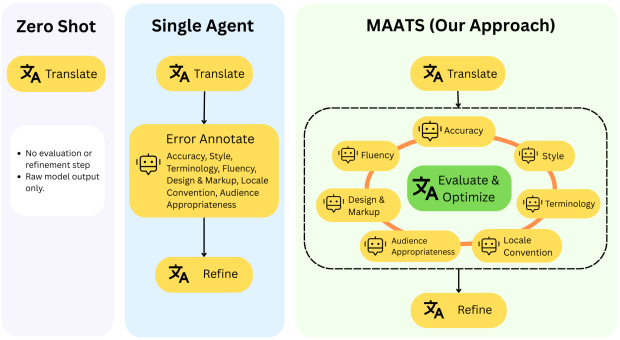

Figure 1: Comparison of Zero-Shot, Single-Agent, and MAATS Pipelines. MAATS assigns MQM dimensions to specialized agents, whose outputs are synthesized by a central agent (Lommel et al. 2024).

This paper introduces MAATS (Multi-Agent Automated Translation System), a novel framework designed explicitly to model a collaborative annotation + refinement process using specialized LLM-based agents. It uses the MQM (Multidimensional Quality Metrics) framework for annotating translation errors across multiple dimensions (Lommel et al. 2024). Each agent is assigned to a distinct MQM dimension to identify translation errors with graded severity (Critical, Major, Minor). Their annotations are then synthesized by a centralized Editor agent, which follows a rule-based priority system. MAATS is evaluated against two baselines: (1) a Single-Agent system that applies MQM-based self-refinement in one step, and (2) a zero-shot direct translation approach with no evaluation or correction guidance.

Our findings are: 1) MAATS consistently outperforms two baselines: zero-shot and single-agent approaches by correcting more critical errors with higher scores on both standard and neural metrics 2) MAATS shows significant neural metric gains for linguistically distant pairs and weaker base LLMs. 3) We conducted an experiment comparing MAATS's annotations with human expert MQM annotations using publicly available data from the WMT MQM Human

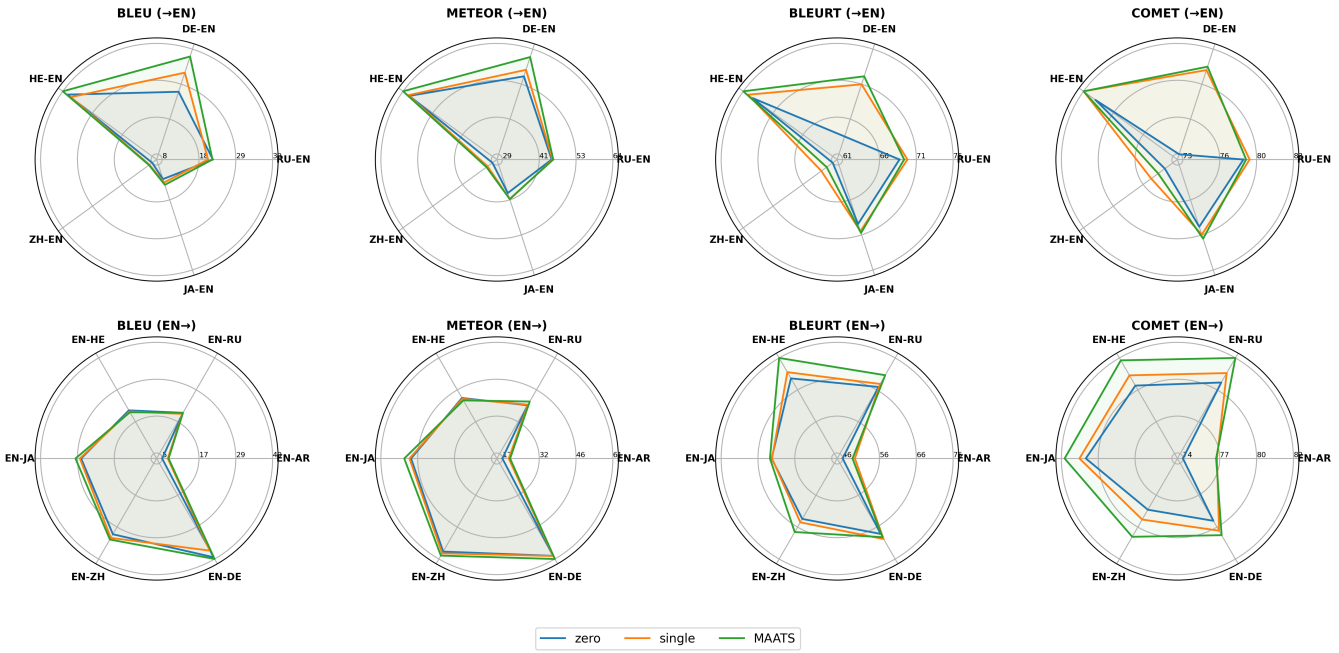

Figure 2: Translation Metrics Comparison for GPT-4o Across Approaches and Language Pairs.

Evaluation dataset (Freitag et al. 2021b). 4) We ran a separate experiment where professional bilingual translators reviewed and ranked translations from MAATS, a single-agent system, and a direct translation baseline. 5) To gain deeper insight into MAATS's translation improvements, a qualitative analysis was conducted through thematic coding and manual examination of its annotation notes.

Background

Recent advances demonstrate that multi-agent frameworks, where multiple LLM agents assume specialized complementary roles, significantly improve outcomes in domains like code generation, planning, and research workflows—examples including SoA, MAGIS, AdaCoder, AutoGen, and GameGPT, where separate agents handle planning, verification, debugging, or design tasks to boost performance over single agents (Ishibashi and Nishimura 2024; Zhu et al. 2025; Pan, Zhang, and Liu 2025; Parmar et al. 2025; Tao et al. 2024). These agents typically interact in structured communication loops, either sequentially or asynchronously. Each agents pass messages, suggestions, or critiques via natural language or structured prompts. They are often mediated by a coordinator or shared memory. This architectural design allows each agent to contribute specialized reasoning from different perspectives, leading to deeper analysis and higher quality outcomes.

In contrast, single-agent translation approaches, such as self-refinement pipelines, rely on a single model critiquing and correcting its own output, which delivers some improvements but often plateau after one iteration and struggles to catch errors due to model bias and limited perspective (Feng et al. 2024). This work situates MAATS in that gap by introducing a collaborative architecture aligned with MQM dimensions. MAATS bridges the gap between human-inspired workflows and black-box LLM translation by combining distributed expertise, interpretable error diagnosis, and structured refinement—yielding more nuanced and faithful translations than both zero-shot and single-agent self-refinement baselines.

Method

MAATS Design

Unlike traditional approaches that combine translation, evaluation, and refinement in one linear step (Feng et al. 2024), MAATS introduces a modular framework where specialized LLM agents collaboratively enhance translation quality through targeted error detection and correction.

Shown in Figure 1, the process begins with a Translator Agent, which generates the initial translation using base LLMs. This output is then evaluated by a set of MQM Evaluator Agents, each aligned with a specific MQM category: Accuracy, Fluency, Locale Convention, Audience Appropriateness, Style, Terminology, and Design & Markup(Lommel et al. 2024). These agents independently annotate translation errors within their domain with critical, major, minor severity levels. All annotations are passed to a centralized Editor Agent, which synthesizes and prioritizes the suggested corrections. The Editor prioritizes resolving critical issues, then addresses less severe ones by severity. We crafted prompts for each agent role. The translator prompt is straightforward (*"Translate the following."*). The evaluator prompt is crucial: for each MQM category, the agent identifies all errors with justification and severity level. We included examples in the prompt to guide annotation format (e.g., showing a

sample error report). The editor agent then consolidates all annotations from the evaluator agents to produce a refined translation. We ensured that the evaluator outputs were included in the editor's context. This modular design avoids iterative feedback loops and ensures efficient integration of evaluations without redundancy or annotation conflicts.

Prompts and Models

MAATS employs distinct prompts for the Translator, specialized MQM Evaluators, and Editor agents. These prompts were designed using a few-shot approach based on real examples adapted from Unbabel's Typology 3.0 (Unbabel 2022). For the Single-Agent baseline, a self-refinement prompt is used (see prompts in supplementary material). Experiments were conducted using three state-of-the-art large LLMs as base models: Claude-3-haiku, Gemini-2.0-flash, and GPT-4o (Anthropic 2024; DeepMind 2025; OpenAI 2024). These models served as the base for all three approaches: zero-shot, single-agent, and MAATS.

One challenge was managing context length, as including the source, translation, and full list of errors can become long. However, with GPT-4 and other models supporting extended context windows, this was manageable for our input size (each input was typically under 1000 tokens). To ensure consistency and reduce randomness during refinement, we kept the model temperature low (between 0 and 0.3).

We also observed that iterating the translation–annotation–refinement loop multiple times did not yield significant improvements. Performance generally plateaued after the first round of refinement, with subsequent iterations resulting in only minor surface-level changes. This aligns with findings that iterative self-refinement often leads to diminishing returns and oscillations in phrasing rather than substantial quality gains (Feng et al. 2024; Xu et al. 2024). Therefore, we limit refinement to a single structured pass to ensure efficiency and stability.

Tasks and Evaluation

To evaluate the MAATS system, translation tasks were conducted bidirectionally between English and six target languages: German (DE), Hebrew (HE), Japanese (JA), Russian (RU), Chinese (ZH), and Arabic (AR) from WMT 2023 (Kocmi et al. 2023) and WMT 2024 (Kocmi et al. 2024). For each direction, we randomly seleted a test set of approximately 200 sentences from the database to balance computational feasibility and alignment with prior MQM studies. These source texts span a range of content, including factual, idiomatic, and technical sentences from news and general domain parallel corpora. The dataset size was chosen to balance feasibility with sufficient coverage for robust metric evaluations and fine-grained error analysis.

To comprehensively evaluate the effectiveness of the MAATS system, we conducted five distinct evaluations, each addressing a different aspect of translation quality.

Automated Metric Comparison. MAATS was evaluated against two baselines using both traditional and neural evaluation metrics. BLEU and METEOR measure surface-level similarity based on n-gram overlap, while BLEURT and COMET are neural metrics designed to assess semantic similarity and fluency (Papineni et al. 2002; Banerjee and Lavie 2005; Sellam, Das, and Parikh 2020; Rei et al. 2020).

Statistical Testing. To determine whether the improvements observed were meaningful, we applied ANOVA and paired bootstrap resampling. These tests confirmed that MAATS's performance gains were statistically significant across most language pairs and model types.

Human-Annotated MQM Evaluation. We validated MAATS's error detection performance by comparing its outputs against human-labeled MQM reference data using confusion matrices (Freitag et al. 2021b).

Human Preference Ranking. We developed a custom web-based interface for this evaluation and recruited professional bilingual translators. Translators were asked to rank translations from the three approaches. We aggregated the rankings using the Borda count method to ensure fair and consistent comparisons(Emerson 2013).

Qualitative Studies. We conducted a qualitative analysis by thematically coding and manually reviewing MAATS's annotation notes to reveal how MAATS identifies and addresses translation issues.

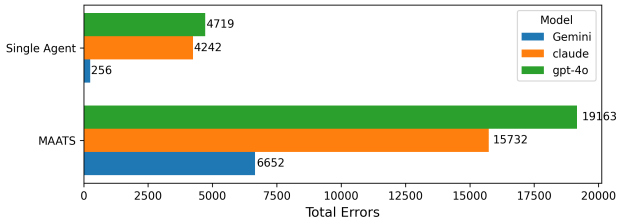

Figure 3: MAATS vs. Single-Agent Total Annotation Counts. GPT found the most errors, followed by Claude and Gemini

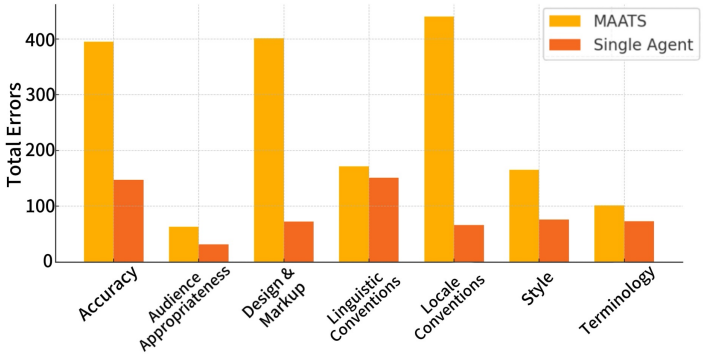

Figure 4: Chinese to English Annotation Analysis Using MAATS and Single Agent

Results

MAATS vs. Baselines: Annotation Scalability and Sensitivity

Our analysis shows that MAATS significantly outperforms the single-agent system in both scale and sensitivity of error

detection. Across all language pairs, MAATS identified a total of 41,547 translation issues, compared to just 9,217 detected by the single-agent baseline—an increase of approximately 450% (see Figure 3). This large difference demonstrates MAATS's superior scalability, which stems from its multi-agent design: each evaluator specializes in a specific MQM dimension, allowing the system to uncover a wider and more nuanced range of translation problems than a single model operating alone.

In terms of sensitivity, MAATS also performs better at identifying detailed and critical errors. For instance, in the Chinese-to-English translation direction (ZH→EN), MAATS detects substantially more issues related to Accuracy, Style, Audience Appropriateness, and Terminology compared to the single-agent approach (see Figure 4). These categories are especially important for preserving meaning and naturalness in translation, and MAATS's ability to capture more of these issues indicates that it is not only finding more errors, but also finding the right kinds of errors that matter for quality. We argue that MAATS provides a more comprehensive and fine-grained view of translation quality. Its ability to scale across languages and models, while remaining sensitive to subtle linguistic and cultural issues, makes it a stronger foundation for downstream translation refinement than conventional single-agent approaches. We conducted evaluation against human MQM annotations to examine MAATS's accuracy in identifying real errors.

Translation Quality Improvements Across Models and Directions

We compare three baselines: direct translation (zero-shot), single-agent refinement (Feng et al. 2024), and our proposed MAATS approach. The results demonstrate that MAATS consistently delivers broader metric gains and higher translation quality, particularly in linguistically distant language directions and under models with lower baseline performance. Figure 2 presents radar plots showing GPT-4o's performance across four evaluation metrics—BLEU, METEOR, BLEURT, and COMET—for both source-to-English and English-to-target directions. MAATS (green line) consistently outperforms both single-agent (orange) and zero-shot (blue) baselines, with the strongest gains observed on neural metrics such as BLEURT and COMET. These metrics are designed to capture semantic adequacy and contextual appropriateness, and they reveal MAATS's ability to produce more fluent, accurate, and contextually aligned translations. Of the LLMs compared, GPT-4o model showed the largest improvements, especially for German-to-English translations, with BLEU rising by +10.6 and COMET by + 8.7, according to Fig. 2. ANOVA test between approaches revealed that MAATS consistently outperformed baselines across most language pairs, with strongest gains on neural metrics like COMET and BLEURT ($p < 0.001$). Improvements were most pronounced in linguistically distant pairs such as EN↔JA, EN↔ZH and EN↔HE and pairwise comparisons (Table 1).

This trend is further confirmed by the metric-level comparison shown in Figure 1, which summarizes translation gains for all three LLMs across 11 language directions. GPT-4o consistently benefits the most from MAATS, achieving improvements on all four metrics in the majority of directions, including EN → DE, DE → EN, EN → JA, JA → EN, and EN → ZH. Gemini also shows notable gains, especially when translating into English, such as in JA → EN, ZH → EN, and RU → EN. This directional asymmetry suggests that MAATS enhances output fluency and semantic alignment when the target language is English, likely due to stronger pretraining coverage and decoder optimization in English. Claude shows more variable results, with fewer 4/4 metric gains, but MAATS still improves its outputs in most cases.

It is also important to distinguish between gains in lexical versus neural metrics. While BLEU and METEOR, which focus on surface-level n-gram overlap, show moderate improvement, BLEURT and COMET reveal more stable and substantial gains. This pattern is especially evident in lower-resource or high-divergence directions, where surface similarity metrics fail to capture the full quality impact of context-aware refinement. For instance, in cases such as GPT-4o on JA → EN or EN → ZH, MAATS leads to marginal BLEU changes but significant BLEURT and COMET improvements. MAATS provides semantic fidelity and naturalness that align more closely with human judgment.

Language Pair	Model	BLEU	METEOR	BLEURT	COMET	Summary
EN_RU	Claude	✓	✗	✓	✗	2/4
	Gemini	✓	✓	✗	✓	3/4
	GPT-4o	✓	✓	✓	✓	4/4
RU_EN	Claude	✗	✗	✗	✓	1/4
	Gemini	✓	✓	✓	✓	4/4
	GPT-4o	✗	✓	✗	✗	1/4
EN_DE	Claude	✗	✗	✓	✓	2/4
	Gemini	✗	✗	✗	✗	0/4
	GPT-4o	✓	✓	✗	✓	3/4
DE_EN	Claude	✗	✗	✗	✓	1/4
	Gemini	✗	✗	✓	✓	2/4
	GPT-4o	✓	✓	✓	✓	4/4
EN_HE	Claude	✗	✗	✓	✓	2/4
	Gemini	✗	✓	✓	✓	3/4
	GPT-4o	✗	✗	✓	✓	2/4
HE_EN	Claude	✓	✓	✗	✓	3/4
	Gemini	✗	✗	✓	✓	2/4
	GPT-4o	✓	✓	✓	✓	4/4
EN_JA	Claude	✓	✓	✗	✓	3/4
	Gemini	✓	✗	✗	✓	2/4
	GPT-4o	✓	✓	✓	✓	4/4
JA_EN	Claude	✓	✓	✗	✓	3/4
	Gemini	✓	✓	✗	✗	2/4
	GPT-4o	✓	✗	✓	✓	3/4
EN_ZH	Claude	✗	✗	✓	✓	2/4
	Gemini	✗	✓	✓	✓	3/4
	GPT-4o	✓	✓	✓	✓	4/4
ZH_EN	Claude	✓	✓	✗	✓	3/4
	Gemini	✓	✗	✓	✓	3/4
	GPT-4o	✓	✓	✗	✗	2/4
EN_AR	Claude	✓	✓	✗	✓	3/4
	Gemini	✓	✓	✓	✗	3/4
	GPT-4o	✓	✓	✗	✗	2/4

Table 1: Translation Gain Comparison (MAATS > Single-Agent > Zero-Shot) Across Language Pairs and Models

MAATS: Higher Sensitivity and Reliability in Quality Assessment

To further validate MAATS's error detection capability and address potential concerns about overestimation, we compared its outputs against human-labeled MQM reference

data using confusion matrix analysis. This evaluation focused on measuring the accuracy of MAATS in identifying real translation errors—particularly its ability to detect true positives and avoid missing critical issues.

The results, shown in the GPT-4o-based confusion matrix (Figure 5), demonstrate that MAATS significantly outperforms the single-agent baseline across all major MQM categories. In the Accuracy dimension alone, MAATS increased true positive detections by 49.7% and reduced false negatives by more than 50%, indicating that it catches many more of the errors that human annotators consider important. Fluency detection improved by a factor of 3.7, while in the Style category, true positives rose by 40% and false negatives dropped by 53.5%. These gains reflect MAATS's improved ability to detect both overt and subtle translation flaws, particularly those that affect readability and tone.

While MAATS does introduce a higher number of false positives compared to the baseline, further manual inspection shows that most of these are low-severity issues (minor phrasing or stylistic inconsistencies), rather than major misjudgments (meaning-changing or comprehension-blocking). This observation is consistent with previous findings in machine translation evaluation, which suggest that many false positives in automated systems reflect debatable but valid concerns about quality (Perrella et al. 2024; Freitag et al. 2021b). In practical terms, this means MAATS errs on the side of caution, surfacing more potentially relevant issues that can be reviewed or filtered downstream. Taken together, these results confirm that MAATS is not simply generating more annotations, it is identifying more meaningful errors. Its high true positive rate and reduced false negative rate indicate a closer alignment with expert human evaluation, and its error detection outputs provide a more actionable basis for improving translation quality.

Human evaluation confirms MAATS preference

To assess whether the improvements observed in our automatic metrics reflect genuine quality gains, we conducted a human evaluation using professional translators. The experiment focused on English-to-Chinese translations from the WMT 2023 test set, comparing outputs generated by MAATS and the two baselines.

A panel of professional bilingual translators (n = 5), each with over three years of industry experience—independently ranked the outputs using the MQM rubric, evaluating the severity and frequency of errors across categories such as Accuracy, Fluency, Style, and Terminology. To avoid bias, all outputs were anonymized and presented in randomized order. The ranking task was implemented via a web-based interface with built-in quality control mechanisms, and we used the Borda count method to aggregate preferences across annotators. As shown in Figure 6, MAATS consistently received higher rankings than both the single-agent and zero-shot systems. Specifically, it achieved win rates of 62.1%, 48.3%, and 49.2% over the single-agent method, and 61.0%, 53.6%, and 51.7% over the zero-shot baseline, across GPT-4o, Claude 3, and Gemini 2.

These results provide strong evidence that MAATS produces translations that are not only more aligned with automatic metrics but also more favorably perceived by human experts. The consistency of these preferences across models and annotators reinforces the validity of our evaluation and supports the conclusion that MAATS delivers more refined and human-preferred outputs.

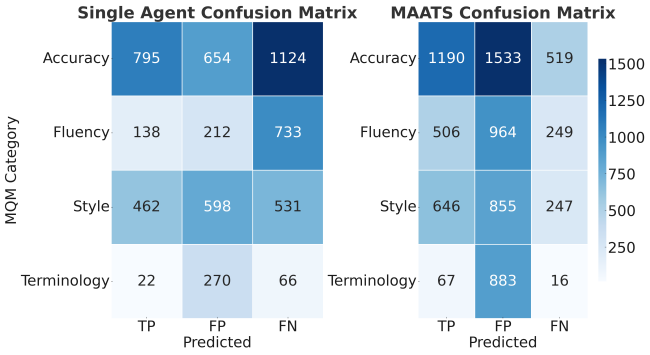

Figure 5: GPT-Based Confusion Matrix

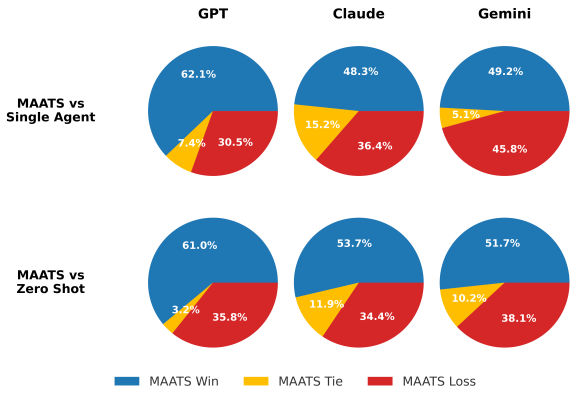

Figure 6: Human Evaluation Results: MAATS vs. Single-Agent and Zero-Shot in EN→ZH Translation

Qualitative analysis reveals MAATS's deeper error understanding

To complement our quantitative evaluations, we conducted a qualitative analysis using three representative case studies that compare MAATS with the single agent baseline. This analysis aimed to examine not just how many errors were detected, but what kinds of improvements MAATS actually contributed to translation quality. We applied a deductive coding approach, guided by predefined MQM categories and refinement types (e.g., semantic accuracy, idiomaticity, contextual adaptation). One researcher performed the initial close reading and annotation, and a second independently reviewed the codes. Disagreements were resolved through discussion, and we achieved high intercoder agreement, indicating strong consistency in theme identification. Through this process, we found that MAATS consistently demonstrates three interrelated strengths: (1) it conducts multi-layered error analysis by identifying both surface-level prob-

	Single Agent	MAATS	Comparison Analysis
Initial Translation	I was looking for a good first car guy car; and I really love the 80s aesthetic. 我在寻找一辆好的第一辆车，而且我真的很喜欢 80 年代的美学。		“好的第一辆车”is literal yet culturally flat; “适合车迷的新手车”carries the enthusiast flavor and sounds far more natural.
Accuracy	[**Major**] Mistranslation – Failed to convey "car guy car" as a car for enthusiasts.	[**Critical**] Omission – "Car guy" omitted. [**Major**] Mistranslation – “好的第一 辆车” does not convey enthusiast meaning.	“好的第一辆车” loses the idea that the buyer is an auto enthusiast. MAATS points out the omission of “car guy” and rewrites it as “适合车迷”, clearly with the fan focus.
Fluency	[**Major**] Punctuation – Semicolon translated as comma.	[**Minor**] Punctuation – Semicolon should be a comma in Chinese (“车； 而且”).	MAATS recommends the proper Chinese form “；” and shows where to place it (“车；而且”)
Style	[**Major**] Awkward – Literal rendering of "car guy car" is unnatural.	[**Minor**] Unnatural Flow – “好的第一辆车” sounds awkward, could be 新手车. [**Minor**] Unnatural Flow – “80 年代的美学” could be “八十年代的复古风格”.	MAATS proposes smoother wording, for example “新手车” for a beginner vehicle and “八十年代的复古风格” for the aesthetic, giving the sentence a natural flow.
Audience Appropriateness	/	[**Minor**] Culture-specific Reference –"Car guy car" may not be understood by Chinese readers; adaptation needed.	MAATS observes that “car guy car” is not a familiar concept for most Chinese readers and adds an explanatory phrase to bridge the cultural gap.
Terminology	/	[**Critical**] Term Not Applied – "Car guy" was not translated per glossary.	MAATS flags “car guy” as a missing term and supplies the approved equivalent “车迷”.
Locale Conventions	/	[**Minor**] Number Format – “第一辆车” doesn’t reflect enthusiast nuance.	MAATS corrects number spacing and the nuance of “第一辆车”, details the Single Agent overlooks.
Final Translation	我在寻找一辆适合车迷的第一辆车；而且我真的很喜欢 80 年代的美学。	我在寻找一辆适合车迷的新手车，而且我真的很喜欢 80 年代风格。	The MAATS version restores the enthusiast angle, uses proper punctuation, and swaps abstract wording for vivid, idiomatic Chinese.
Reference	我正在找一辆适合汽车迷的新手车，而且我特别喜欢八十年代的复古风格。		

Table 2: MAATS Restores Enthusiast Nuance and Cultural Naturalness in the “Car Guy Car” Translation

lems and deeper semantic mismatches; (2) it detects omissions and subtle issues across multiple MQM dimensions, resulting in broader and more precise diagnostic coverage; and (3) it improves contextual alignment by adjusting tone, refining idiomatic choices, and adapting cultural references to better match the expectations of the target audience. These findings reinforce the conclusion that MAATS offers more than broader coverage—it enables more sophisticated, human-like translation refinement.

Case Study 1: “Car Guy Car” – Multi-Layered Error Analysis for Cultural Misalignment

The first case demonstrates MAATS’s stronger ability to identify the root causes of translation issues and classify them with precision. In Table 2, the source phrase “car guy car” refers to a vehicle suited for someone passionate about cars—an enthusiast’s first ride. The Single Agent translated it literally as “a good first car,” which completely misses the nuance that the buyer is an auto enthusiast. It categorized this mistake only as a style issue, vaguely labeling the result as “awkward,” without recognizing the deeper cultural or contextual mismatch.

In contrast, MAATS flagged multiple overlapping problems. It identified the loss of meaning as a critical terminology error, pointed out that the concept of “car guy” may not be familiar to the target audience and requires cultural adaptation, and also noted formatting issues in how the phrase “first car” was presented. In the final output, MAATS replaced the flat translation with a clearer and more expressive version that adds context—essentially rephrasing it as “a beginner car suited for enthusiasts.” This preserved the intent behind “car guy car,” which the Single Agent had overlooked. Additionally, MAATS improved the translation of the phrase “80s aesthetic.” While the Single Agent used a generic term for aesthetics, MAATS chose language that conveyed a more vivid sense of retro style, aligning better with how a native speaker would describe design preferences from the 1980s.

Case Study 2: “Book the Fare” – Broad-Spectrum Error Detection Through Multi-Dimensional Omissions

The second case, shown in Table 3, highlights MAATS’s ability to detect and resolve translation issues across multiple quality dimensions, particularly in cases involving layered omissions. The original phrase, “We provide links to where you can book the fare,” carries two critical pieces of meaning: the idea of booking a flight and the directionality implied by “where.” Only MAATS systematically diagnosed the full range of errors that contributed to this failure. The Single Agent missed the mark entirely. It offered no annotations on fluency, style, audience, or terminology, and flagged only the omission of “to where.” This narrow diagnosis resulted in a literal and culturally flat phrase that misrepresented the intent of the original sentence.

MAATS, by contrast, engaged in broad-spectrum error detection. It not only identified the missing phrase “to where” as a critical omission under Accuracy, but also flagged the literal rendering of “book the fare” as a major mistranslation that fails to reflect real-world usage. Beyond that, MAATS annotated the output as awkward (Style – Minor), culturally unnatural (Audience – Minor), and incomplete in applying the required term (Terminology – Critical). These overlapping annotations reflect a multidimensional

	Single Agent	MAATS	Comparison Analysis
Initial Translation	We provide links to where you can book the fare. 我们提供可以**预订票价**的链接。		**票价 (Fare)** is literally correct but culturally unnatural; **"book a flight (机票)"** is more idiomatic in Chinese.
Accuracy	[**Major**] Omission – The phrase "to where" is omitted, which specifies the destination or platform.	[**Critical**] Omission – "to where" is omitted from the translation. [**Major**] Mistranslation – "预订票价" is literal but doesn't capture the contextual meaning of providing links for booking.	MAATS flagged both omission and mistranslation; Single only noted omission.
Style	/	[**Minor**] Unnatural Flow – "可以预订票价的链接" is correct but slightly awkward. More natural: "可以预订机票的链接" or "票价预订链接".	MAATS noted awkward phrasing and offered natural alternatives; Single missed it.
Audience	/	[**Minor**] Culture-specific Reference – "预订票价" is not natural in Chinese. Suggested alternatives: "预订机票" or "预订航班".	MAATS adapted for cultural clarity; Single had no awareness of reader expectations.
Terminology	/	[**Critical**] Term Not Applied – "Fare" should be highlighted per glossary but wasn't.	MAATS attempted to highlight "fare" as a term; Single ignored it.
Final Translation	我们提供链接，你可以通过这些链接**预订票价(fare)**。	我们提供可以**预订机票(flight)**的链接。	MAATS localized to '机票 flight'; Single kept '票价 fare'.
Reference	我们会给您提供**预订机票(flight)**的链接。		

Table 3: MAATS Corrects Cultural and Contextual Errors in "Book the Fare. By flagging Accuracy, Style, Audience, and Terminology, a more natural and contextually appropriate term was revised.

perspective that allowed MAATS to reconstruct the meaning more faithfully in the final translation, replacing "book the fare" with a clearer, idiomatic expression equivalent to "book a flight." While MAATS's terminology flag on "fare" was partially imprecise, it highlighted the term but did not directly resolve its misinterpretation. The system still arrived at the right fix. This shows that MAATS's multi-layered framework can serve as a self-correcting mechanism.

Discussion

Our results show that decomposing translation quality into specialized, collaborating agents yields benefits that extend beyond incremental metric gains. First, MAATS' large advantage in semantic metrics and human rankings suggests that current "all-purpose" LLMs under-serve high-stakes translation requirements; distributing expertise across MQM dimensions lets each agent focus on a narrower decision space, mirroring the translator–editor division found in professional workflows. For the machine-translation community, this supports a shift from single-model optimization toward orchestrated micro-tasks in which accountability and interpretability are built in at the architectural level. Second, the disproportionate improvements on linguistically distant pairs and on weaker base models indicate that modular collaboration can compensate for limited pre-training coverage. This implies that multi-agent refinement may provide a practical path for elevating mid-tier or domain-specific LLMs when access to top-tier models is restricted—an important consideration for low-resource languages and on-premise deployments. Third, MAATS' higher true-positive rate + moderate false-positive inflation reveals a quality-control trade-off: cautious over-reporting of minor issues versus the risk of critical omissions in single-agent self-review. Because MQM already encodes error severity, pipeline designers can selectively surface only Major/Critical flags, reducing reviewer load while retaining MAATS' diagnostic depth. For the broader multi-agent-systems field, our ablation shows that one well-structured pass delivered most of the gains; repeated negotiation loops offered diminishing returns. This finding aligns with recent work in planning and code generation and motivates research on adaptive agent activation, where only the dimensions most likely to add value are invoked.

Conclusion

In conclusion, MAATS bridges the gap between black-box LLMs applied to MT and human translation workflows by using interpretable MQM dimensions and modular roles. It shifts the focus from superficial fluency toward deeper semantic fidelity and contextual alignment, producing outputs that better reflect human standards. Its distributed architecture not only boosts translation performance but also introduces robustness and modularity—laying the groundwork for broader applications of multi-agent collaborative AI in structured NLP tasks. Specifically, MAATS excels at nuanced error detection, contextual refinement, and culturally adaptive translation, achieving substantial improvements in semantic accuracy and fluency as validated by both automated and human evaluations. While MAATS showcases clear strengths, some limitations remain. The system's multi-agent architecture inherently introduces increased computational overhead due to the involvement of multiple specialized agents and subsequent annotation integration. Additionally, false positives are more frequent compared to single-agent methods, although typically confined to minor stylistic issues rather than critical translation errors. Future research could explore extending the multi-agent paradigm beyond translation into other structured NLP tasks, such as multilingual summarization, sentiment analysis, or dialogue generation, to validate the generalizability and effectiveness of this modular collaborative approach. Further studies could also investigate methods to reduce computational complexity through selective agent activation or dynamic prioritization strategies, thereby optimizing resource efficiency while maintaining quality gains.

References

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Accessed: 2025-05-18.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Ann Arbor, Michigan: Association for Computational Linguistics.

DeepMind, G. 2025. Gemini 2.0 Flash | Generative AI on Vertex AI. Accessed: 2025-05-18.

Emerson, P. 2013. The original Borda count and partial voting. *Social Choice and Welfare*, 40(2): 353–358.

Feng, Z.; Zhang, Y.; Li, H.; Wu, B.; Liao, J.; Liu, W.; Lang, J.; Feng, Y.; Wu, J.; and Liu, Z. 2024. TEaR: Improving LLM-based Machine Translation with Systematic Self-Refinement. arXiv:2402.16379.

Freitag, M.; Foster, G.; Grangier, D.; Ratnakar, V.; Tan, Q.; and Macherey, W. 2021a. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9: 1460–1474.

Freitag, M.; Foster, G.; Grangier, D.; Ratnakar, V.; Tan, Q.; and Macherey, W. 2021b. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. arXiv:2104.14478.

Ishibashi, Y.; and Nishimura, Y. 2024. Self-Organized Agents: A LLM Multi-Agent Framework toward Ultra Large-Scale Code Generation and Optimization. arXiv:2404.02183.

Jiao, W.; Wang, W.; tse Huang, J.; Wang, X.; Shi, S.; and Tu, Z. 2023. Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. arXiv:2301.08745.

Kamoi, R.; Das, S. S. S.; Lou, R.; Ahn, J. J.; Zhao, Y.; Lu, X.; Zhang, N.; Zhang, Y.; Zhang, R. H.; Vummanthala, S. R.; Dave, S.; Qin, S.; Cohan, A.; Yin, W.; and Zhang, R. 2024. Evaluating LLMs at Detecting Errors in LLM Responses. arXiv:2404.03602.

Kocmi, T.; Avramidis, E.; Bawden, R.; Bojar, O.; Dvorkovich, A.; Federmann, C.; Fishel, M.; Freitag, M.; Gowda, T.; Grundkiewicz, R.; Haddow, B.; Karpinska, M.; Koehn, P.; Marie, B.; Monz, C.; Murray, K.; Nagata, M.; Popel, M.; Popović, M.; Shmatova, M.; Steingrímsson, S.; and Zouhar, V. 2024. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In *Proceedings of the Ninth Conference on Machine Translation*, 1–46. Miami, Florida, USA: Association for Computational Linguistics.

Kocmi, T.; Avramidis, E.; Bawden, R.; Bojar, O.; Dvorkovich, A.; Federmann, C.; Fishel, M.; Freitag, M.; Gowda, T.; Grundkiewicz, R.; Haddow, B.; Koehn, P.; Marie, B.; Monz, C.; Morishita, M.; Murray, K.; Nagata, M.; Nakazawa, T.; Popel, M.; Popović, M.; Shmatova, M.; and Suzuki, J. 2023. Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet. In *Proceedings of the Eighth Conference on Machine Translation*, 1–42. Singapore: Association for Computational Linguistics.

Lommel, A.; Gladkoff, S.; Melby, A.; Wright, S. E.; Strandvik, I.; Gasova, K.; Vaasa, A.; Benzo, A.; Sparano, R. M.; Foresi, M.; Innis, J.; Han, L.; and Nenadic, G. 2024. The Multi-Range Theory of Translation Quality Measurement: MQM scoring models and Statistical Quality Control. arXiv:2405.16969.

Manakhimova, S.; Avramidis, E.; Macketanz, V.; Lapshinova-Koltunski, E.; Bagdasarov, S.; and Möller, S. 2023. Linguistically Motivated Evaluation of the 2023 State-of-the-art Machine Translation: Can ChatGPT Outperform NMT? In Koehn, P.; Haddow, B.; Kocmi, T.; and Monz, C., eds., *Proceedings of the Eighth Conference on Machine Translation*, 224–245. Singapore: Association for Computational Linguistics.

OpenAI. 2024. GPT-4o System Card. Accessed: 2025-05-18.

Pan, R.; Zhang, H.; and Liu, C. 2025. CodeCoR: An LLM-Based Self-Reflective Multi-Agent Framework for Code Generation. arXiv:2501.07811.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318.

Parmar, M.; Liu, X.; Goyal, P.; Chen, Y.; Le, L.; Mishra, S.; Mobahi, H.; Gu, J.; Wang, Z.; Nakhost, H.; Baral, C.; Lee, C.-Y.; Pfister, T.; and Palangi, H. 2025. PlanGEN: A Multi-Agent Framework for Generating Planning and Reasoning Trajectories for Complex Problem Solving. arXiv:2502.16111.

Peng, K.; Ding, L.; Zhong, Q.; Shen, L.; Liu, X.; Zhang, M.; Ouyang, Y.; and Tao, D. 2023. Towards Making the Most of ChatGPT for Machine Translation. arXiv:2303.13780.

Perrella, S.; Proietti, L.; Cabot, P.-L. H.; Barba, E.; and Navigli, R. 2024. Beyond Correlation: Interpretable Evaluation of Machine Translation Metrics. arXiv:2410.05183.

Qian, M.; and Kong, C. 2024. Enabling Human-Centered Machine Translation Using Concept-Based Large Language Model Prompting and Translation Memory. In *Artificial Intelligence in HCI: 5th International Conference, AI-HCI 2024, Held as Part of the 26th HCI International Conference, HCII 2024, Washington, DC, USA, June 29–July 4, 2024, Proceedings, Part III*, 118–134. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-031-60614-4.

Rei, R.; Stewart, C.; Farinha, A. C.; and Lavie, A. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702.

Sellam, T.; Das, D.; and Parikh, A. P. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7881–7892.

Tao, W.; Zhou, Y.; Wang, Y.; Zhang, W.; Zhang, H.; and Cheng, Y. 2024. MAGIS: LLM-Based Multi-Agent Framework for GitHub Issue Resolution. arXiv:2403.17927.

Unbabel. 2022. Annotation Guidelines: Typology 3.0. Online resource. Accessed2025-05-19.

Wu, M.; Yuan, Y.; Haffari, G.; and Wang, L. 2024. (Perhaps) Beyond Human Translation: Harnessing Multi-Agent Collaboration for Translating Ultra-Long Literary Texts. arXiv:2405.11804.

Xu, W.; Zhu, G.; Zhao, X.; Pan, L.; Li, L.; and Wang, W. Y. 2024. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. arXiv:2402.11436.

Yan, J.; Yan, P.; Chen, Y.; Li, J.; Zhu, X.; and Zhang, Y. 2024. Benchmarking GPT-4 against Human Translators: A Comprehensive Evaluation Across Languages, Domains, and Expertise Levels. arXiv:2411.13775.

Yan, R.; Gao, M.; Pavlick, E.; and Callison-Burch, C. 2014. Are Two Heads Better than One? Crowdsourced Translation via a Two-Step Collaboration of Non-Professional Translators and Editors. In Toutanova, K.; and Wu, H., eds., *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1134–1144. Baltimore, Maryland: Association for Computational Linguistics.

Zhu, Y.; Liu, C.; He, X.; Ren, X.; Liu, Z.; Pan, R.; and Zhang, H. 2025. AdaCoder: An Adaptive Planning and Multi-Agent Framework for Function-Level Code Generation. arXiv:2504.04220.