Multi-Agent Translation Team (MATT): Enhancing Low-Resource Language Translation through Multi-Agent Workflow

Anishka Peter[1], Mai Dang[1], Michael Liu[1], Nibhrat Lohia[1], Joaquin Dominguez[2]
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
[2] New Haven, CT 06511 USA
{apeter, maid, haitiel, nlohia}@mail.smu.edu
joaquin.dominguez@proton.me

Abstract. Like humans, large language models (LLMs) benefit from revision and refinement, especially for complex tasks requiring critical thinking. Inspired by human collaborative problem-solving, this study introduces a novel multi-agent workflow designed to enhance LLM translations from English to low-resource languages. Multi-Agent Translation Team (MATT) involves the collaboration of agents that are assigned specific roles, such as translator, evaluation coordinator, and various levels of editing, to refine the initial translation into the most desired version possible. The agents work collaboratively in an iterative loop until the translation loss meets a satisfactory threshold. It stands out from other multi-agent workflows by combining the strengths of LLMs and Google Translate (GT) to achieve higher translation quality. This approach shows promise in translating short sentences and long chunks from English to languages such as Vietnamese, Hindi, and Malayalam.

1 Introduction

Language plays a vital role in every culture as it is the means through which knowledge is shared and passed on. As a result, quality translation is essential to effectively spread information between cultures that do not speak the same language. Even though the world has become increasingly globalized with English as the dominant language, only about 15% of the world's population speaks English which leaves a substantial portion of written content inaccessible to most of the world (Gillan, 2020). Among Asian countries, Vietnam is one of the top ten countries that have the most foreign investments from English spoken countries, but they have a limited population of people proficient in English (The Demographics, Linguistics, and Importance of Learning English in Vietnam, 2023).

Not only do non-English speakers miss out on discoveries written in English, but English speakers also miss out on access to information from other cultures. Limiting access to information for some groups of people hinders the ability of information to benefit all population groups equally. For example, many medical journals are written only in English, and without proper translation abilities, other non-English dominant cultures may miss out on life-saving medical advancements. Accurate translation of all

languages is essential to enable cultures to collaborate in expanding global ideas and share discoveries that enrich humanity.

Language is not merely a tool for basic communication. It encapsulates cultural, social, and historical nuances which are crucial for comprehension. Understanding these nuances enriches the appreciation of the depth and complexity embedded in linguistic expressions, and it is critical when trying to represent the original words in the targeted translation language. Furthermore, certain concepts exist in one language that are not in the other. For example, in Vietnamese, the proper personal pronouns depend on the speakers social standing, and attitude in the conversation which is not a consideration in English (Carl, 2023). Implicit connotations of words, such as these, make translation tasks very subjective and dependent on a skilled translator to navigate the cultural and linguistic meanings of words to produce accurate translations.

Since translation depends on the implicit connotation of words rather than strict rules, the task of translation is difficult for humans and even more so for machines. Since translation is important in creating a more interconnected world, in 2018 the US Bureau of Labor and Statistics predicted that the translation industry would grow at a rate of 46% during the period 2012-2022 (Manasse, 2018). Compared to machines, human translators can be slower, cost more, and risk confidentiality (Manasse, 2018). The increasing need for translators and the limitations of human translators continue to increase the need for accurate translation tools.

Currently, many machine translation tools rely on neural machine translation architecture, most notably Google Translate (GT) which has reports showing up to 94% accuracy depending on the language ("How Accurate Is Google Translate?", n.d.). This architecture has proven to improve translation but is still challenged by large vocabularies, rare words, and low-resource languages, such as Vietnamese, Hindi, and Malayalam (Mohamed et al, 2021). Ironically, countries like Vietnam and India, which could greatly benefit from reliable translation tools, face a shortage of effective machine translation solutions due to the complexities and unique differences in their semantic meanings and language structures. However, Natural Language Processing (NLP) has been revolutionized through the emergence of Large Language Models (LLMs) with their ability to generate coherent text, decompose complex queries, and perform a wide range of linguistic tasks. Recent studies have shown that larger LLMs have strong translation abilities with a promising future but still perform worse than current translation standards such as GT (Zhu et al., 2023). Furthermore, the translation abilities of these LLMs significantly decrease for low-resource languages (Zhu et al., 2023).

One proposed approach to addressing LLMs' challenges with translation is the development of multi-agent workflows. These workflows consist of a network of autonomous agents—independent LLMs each with specified tasks—that work collaboratively to tackle complex problems. This model draws inspiration from human problem-solving teams, where the synergy of diverse skills and perspectives leads to more effective solutions than individual efforts alone. In multi-agent workflows, each agent contributes their expertise, receives and provides feedback, and collectively drives towards solving intricate problems. Integrating such workflows with LLMs creates the potential to improve translation quality (Wyndham, 2024).

The application of a multi-agent workflow introduces a more dynamic and adaptable approach to problem-solving, mimicking the collaborative nature of human

teams. This shift is pivotal for creating systems that can evolve, adapt, and respond to new challenges and environments in real time. It is reasonable to assume such systems are more robust and capable of handling a wider range of tasks, making them more valuable in diverse applications.

Furthermore, integrating multi-agent workflow with LLMs could set a precedent for future research and development. It underscores the importance of collaboration and specialization, paving the way for more sophisticated and versatile architecture. Investigating this approach addresses current challenges and contributes to the long-term goal of developing systems that exhibit human-like problem-solving capabilities.

The concept of multi-agent workflows is at the forefront of contemporary research. Various scholars and researchers have explored how multiple autonomous agents can work together to enhance problem-solving abilities (Wu et al., 2024). For instance, in April 2024, Andrew Ng, a prominent generative intelligence (GenAI) researcher, highlighted the effectiveness of multi-agent workflows in a keynote address. He discussed how coordinating efforts among specialized LLMs allows for optimized performance in problem-solving tasks. Ng's insights reflect a broader trend in GenAI research that emphasizes the value of collaborative systems. Through agents' contribution of their unique insights and strategies, the overall efficiency and accuracy of the systems are enhanced.

Recently, researchers have created multi-agent translation workflows that translate various high-resource languages and have found that the models perform poorly when looking at traditional scoring metrics such as BLEU scores. Despite mediocre performance according to this quantitative metric, they have promising results from the collaborative nature of multi-agent systems (Wu et al., 2024). This research aims to use a multi-agent workflow to address the challenges of understanding and translating the complex nuances of English that LLMs currently face when translating English to Vietnamese, Hindi, and Malayalam. Vietnamese and Hindi are national languages of Vietnam and India respectively, and Malayalam is a small regional language within India. These languages are considered low-resource which means there is not a lot of training data available for LLMs. By designing and implementing a network of autonomous agents with specialized roles, this research aims to enhance the models' capability to understand and translate the semantic meaning of phrases for low-resource languages. Each agent will focus on one aspect of translation, contributing to a collaborative system that mirrors human teamwork. This approach is expected to improve the overall performance of LLMs in translation tasks, making them more effective and versatile.

2 Literature Review

The literature review focuses on four principal areas: machine translation, LLM models, the current machine translation workflows, and multi-agent workflows in various domains.

2.1 Translation History

Translation is said to have originated during the Mesopotamian era and was prevalent in ancient societies because of the need to communicate between ancient kingdoms. The first significant translation was of the Hebrew Bible in the 3rd century and the desire to spread religion increased the need to translate text into multiple languages in ancient societies (*Brief History of Translation: Everything You Need to Know*, n.d.). While religious needs were a major contributor for the need of translation services, the number of reasons for translation has greatly expanded.

In the past, human translation was the only method available for translating languages. However, today, there are numerous tools to assist human translators, as well as fully automated computer-based translation systems. There are currently around 7000 active languages, however, only a fraction of those languages is actively being translated for the masses (*Over 7000 Languages Are Spoken in the World Today, but Not Many Are Represented Online - Consumers International*, 2018). Currently there are 243 total languages supported by the largest translation tool GT, and 110 of those are new languages added in 2024 (Whitney, 2024).

However, the translation quality varies widely between languages. The accuracy of these languages is between 55% and 94% where more high-resource languages such as Spanish, English, and French have higher accuracy than low-resource languages such as Armenian (Schoening, 2023). A high-resource language has a significant amount of training data available, in fact about 75% of the content available online is in high-resource languages which make up only 6 of the 7000 languages in the world (Petrosyan, 2024). The large volume of linguistic resources available for them enables the development of machine translation resources in these languages (Mirela, 2024). Despite the difficulty of training models on low-resource languages, there are continued initiatives to improve translation quality and quantity. For example, Google has announced their 1000 languages initiative to use GenAI models to increase the number of translation languages supported (Whitney, 2024).

2.2 Machine Translation

Machine Translation (MT) is a process of using computing power to translate a text or corpus of text into another target language without human involvement (Yuan et al, 2023). The long history of MT goes back to René Descartes who in the 17th century, proposed a universal language that conveyed the same meaning across different languages through shared symbols (Yang et al., 2020). Significant advancements in MT only began in the 1950s, notably with Yehoshua Bar-Hillel organizing the first International Conference on MT in 1952. Since then, MT has evolved alongside the modernization of computers and machines.

The advent of Statistical Machine Translation (SMT) presented a breakthrough in the 1990s. SMT models use statistical methods to analyze bilingual text corpora and identify patterns that can be used to translate text. These models rely on large amounts of parallel text data (texts that are translations of each other) to build statistical models of language. (Stahlberg, 2020). SMT became the mainstream technology for MT for about 20 years, with successful applications in various commercial translation services like GT and Baidu Translate (Yang et al., 2020).

The introduction of Neural Machine Translation (NMT) in the mid-2010s represented a paradigm shift in the field of MT. NMT models use deep neural networks to learn the translation process, employing an end-to-end approach where a single neural network directly transforms the source sentence into the target language. One of the earliest successful NMT models was the encoder-decoder model with attention mechanisms, introduced by Bahdanau et al. in 2015. This model improved translation quality by allowing the network to focus on various parts of the source sentence during translation (Stahlberg, 2020).

The adoption of NMT has led to substantial advancements in translation accuracy and fluency. Modern NMT systems, such as the Transformer model introduced by Vaswani et al. in 2017, have further enhanced performance by using self-attention mechanisms to capture long-range dependencies in text more efficiently. NMT is now widely used in commercial translation services, providing more natural and contextually appropriate translations (Yang et al., 2020).

2.3. LLMs and Translation Tasks

There is a growing trend towards incorporating pre-trained LLMs for machine translation (Yang et al., 2019; Brown et al., 2020). Due to their superior natural language understanding and generation and extensive pretraining, LLMs are expected to enhance MT significantly (Lewis et al., 2019).

In a preliminary study on GPT-4's translation abilities, Jiao et al. (2023a) examined various aspects, including its translation prompts, multilingual translation, and robustness. They found that GPT-4 has a competitive performance when compared with popular translation products, such as GT and DeepL, for high-resource European languages. However, the paper also highlighted that ChatGPT's performance lags significantly behind other MT commercial systems when handling low-resource languages.

From another perspective, GPT-4 also shows competitive results in spoken language translation but underperforms on biomedical abstracts or Reddit comments compared to commercial ML systems (Jiao et al., 2023b). These exemplify one of GPT's most well-known limitations, general domain handling. It often performs poorly in domains where it has not been extensively pre-trained. GPT could perform more efficiently if only domain-specific information was included in the translation prompt. Other LLMs also display the same limitations when translating a text in which no further domain knowledge is provided (Yuan et al., 2023).

2.4 NMT and LLMs

While both NMT and LLM translation have significantly transformed the field of MT, they have many limitations when used alone. As a result, recent research focuses on combining these technologies as one method of improving translation quality (Oh et al., 2023, Clinchant et al., 2019, Santy et al., 2019, Jiao et al., 2023, Lyu et al, 2023). One problem that traditional MT faces is producing translations that match a specific

genre or style because of the lack of extensive training corpora in various languages. However, the zero-shot capabilities of LLMs have made it possible to achieve stylized MT. Toshevska and Gievska (2022) and Wang et al. (2022) proposed that NMT could better handle tasks requiring stylistic variations, such as formal or informal language when used in conjunction with LLMs. Not only do LLMs allow for better generation of stylistic machine translation, but their natural language generation abilities offer a promising solution to the lack of extensive training corpora. LLMs can be used for prompt-based data augmentation using LLMs to generate synthetic parallel data, reducing the time and cost associated with data creation, especially for low-resource languages (Oh et al., 2023).

Experiments have shown that appropriate prompts can enhance the diversity and quality of training data, leading to improved translation performance (Oh et al., 2023). Additionally, training the LLM on the source language and fine-tuning the LLM helps the NMT provide a better translation (Clinchant et al., 2019). NMT has also been improved by allowing users to correct or refine automatic translations and providing feedback during the translation process. Santy et al. (2019) and Jiao et al. (2023) emphasized that the integration of MT systems based on LLMs with interactive user interfaces enhances translation accuracy and fluency, especially when dealing with ambiguous source language or limited domain knowledge. Integrating new LLMs into existing NMT systems brings the superior NLP skills of LLMs which have great potential for improving MT systems. Customized prompts, stylized MT, interactive translation interfaces, and prompt-based data augmentation are some of the innovative approaches being explored. Despite the challenges, such as defining stylistic variations and incorporating user feedback, advancements in this field promise to enhance the accuracy, efficiency, and robustness of MT systems.

2.5 Agents

Leveraging the abilities of LLMs remains an active area of research with various approaches under exploration. The development of multi-agent systems is one approach to effectively harness the power of LLMs. An agent is a software program that can autonomously perform various tasks and meet human-determined goals. Agents powered by LLMs can perform tasks that require complex reasoning by breaking the tasks into smaller parts and acting on them (Varshney, 2023 & Guo et al., 2024). Agents can also work together in a multi-agent system which mimics human teams where agents have diverse profiles and interact with each other to solve dynamic and complex tasks. These multi-agent systems have been used in various applications from software development to multi-robot collaboration and conducting science experiments. Extensive research has shown their ability to simulate real-world situations such as societal, gaming, and economic simulations (Guo et al., 2024).

Andrew Ng outlined important design elements of multi-agent systems that enhance the agents' ability to accomplish these complex tasks. Agents can be given multiple tools so that the LLM can leverage the appropriate one to complete the task. Furthermore, implementation of reflection agents has exhibited potential to improve multi-agent solutions. Reflection agents can self-reflect on the initial response and

identify and improve on errors (Ng, 2024b). Madaan et al. (2023) introduced the concept of interactive refinement with self-feedback and found that the self-reflection mechanism significantly enhanced the performance of all the LLMs used in tasks such as text generation and question answering. Furthermore, integrating verbal reinforcement learning either from humans or other agents has been seen to create more accurate and contextually appropriate responses from LLMs (Shinn et al., 2023). This method of reflection improves the quality of responses. However, it still depends on the LLMs ability of quality self-evaluation which is not guaranteed (Liang et al., 2023).

2.6 Translation Multi-Agent Workflows

Because of the limitations of extensive parallel corpora, translation tasks are often regarded as one of the most difficult tasks in NLP (Wu et al., 2024). Increasing the amount of training data of LLMs is extremely expensive, however, the introduction of multi-agent systems allows for collaboration between different LLM agents to generate better responses. Looking at translation at the sentence level, the response quality improved when multiple LLMs with distinct personas were used, compared to relying on a single LLM. (Wang et al., 2024).

Translation_Agents. One of the pioneering multi-agent workflows for translation is Andrew Ng's "Translation_Agents." While not a fully developed software framework, it serves as a foundational concept for multi-agent-based translation systems. The workflow emphasizes iterative refinement, where translations are generated, reflected upon, and improved through collaboration among agents. These agents, powered entirely by LLMs, work together in a structured process to enhance translation quality, showcasing the potential of LLMs in collaborative tasks.

TRANSAGENTS. Another framework used a debate and judge to encourage divergent thinking of LLMs and reduced biases and enhanced the nuanced accuracy of translations (Liang et al., 2023). For translation of longer texts, Wu et. al (2024), implemented a virtual multi-agent translation company, TRANSAGENTS. Across all genres the translations score poorly using traditional translation evaluation metrics such as BLEU scores, however the human and LLM evaluators preferred the translation from TRANSAGENTS compared to reference and GPT-4 translations (Wu et al., 2024).

2.7 Translation Evaluation

Human Evaluation. The final step in translation is the evaluation of the translation quality. Translation is subjective, with its appropriateness varying based on individual judgment. As a result, human evaluation remains the gold standard for translation quality assessment. However, it is time-consuming, expensive, and subject to inter-annotator variability so using some metric is necessary. The subjective nature makes it

difficult for traditional metrics to determine the quality of translations uniformly across languages.

BLEU Scores. BLEU (Bilingual Evaluation Understudy) has been widely used to evaluate MT, but it often fails to capture nuanced linguistic and contextual elements (Papineni et al., 2002; Denkowski & Lavie, 2011). Translation evaluation has become more difficult with the introduction of multi-agent translation frameworks because the alignment and coherence of the final translation need to be assessed holistically, which traditional metrics are ill-equipped to handle. Overall, LLM translations and multi-agent translations perform worse according to BLEU scores compared to human evaluations (Zhang et al., 2020, Clinchant et. al, 2019, Wu et. al, 2024).

LLMs Evaluation. The abilities of LLMs have introduced more sophisticated scoring techniques by leveraging pre-trained language models like BERT to evaluate translations. Kocmi & Federmann (2023) found that GPT-4 accurately estimated translation quality without requiring the reference of human translations (Wu et. al, 2024). BERT showed an improved correlation with human judgment but still struggled with long-range dependencies and stylistic variations (Zhang et al., 2020). However, studies have shown that an LLM evaluator tends to favor the translation output from an LLM over the output from other translation systems, which may influence the accuracy of LLM translation evaluations (Lyu et. al, 2023). Some strategies introduced specifically for multi-agent translation evaluations are Monolingual Human Preference (MLP) and Bilingual LLM Preference (BLP) (Wu et al., 2024). Both MLP and BLP favored multi-agent translations despite lower BLEU scores (Wu et al., 2024). Recent research has introduced promising approaches in evaluating translations from multi-agent collaboration. However, there is not a common method, but the best evaluation method depends on the text and the translation framework used.

3 Methods

This paper introduces the Multi-Agent Translation Team (MATT), an approach to enhancing MT through a multi-agent workflow inspired by Andrew Ng's "translation_agents" framework. "Translation_agents" serves as the Baseline model, both as a benchmark for comparison and as the foundation for MATT. The Baseline architecture has an initial translation generated by an LLM and then improves the translation one time based on feedback from a reflection agent (Appendix A). Building on this, MATT introduces enhanced agent roles, advanced evaluation mechanisms, iterative refinement, and the use of GT to address challenges in translation, particularly for low-resource languages.

MATT integrates GPT-4o and Llama 3.1 across five key components: a multi-agent workflow layer, tool utilization, a voting mechanism, an evaluation process, and an iterative improvement loop (Appendix E). To address initial translation errors, MATT incorporates the GT API. Performance is evaluated through bilingual human

assessments, *sp*-BLEU scores, and LLM-based evaluations. The study focuses on translating English into Vietnamese, Hindi, and Malayalam, highlighting MATT's effectiveness in improving the translations' accuracy, fluency, style, and terminology for low-resource languages.

3.1 Multi-agent Translation Team (MATT)

The core methodology of this research is the multi-agent workflow layer, which is comprised of two distinct parts designed to iteratively enhance translation quality. The general architecture is shown below in Fig. 1.

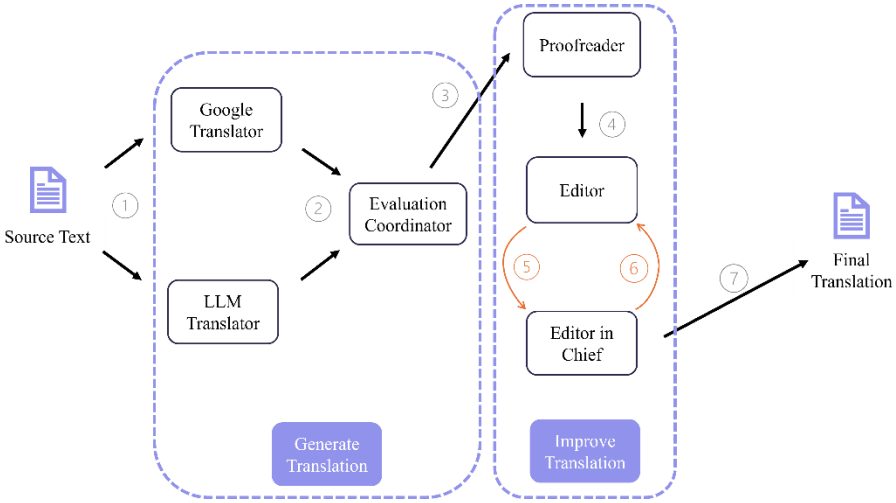

Fig. 1. The plot shows MATT's architecture.

Generating Initial Translation. The first layer involves generation of initial translations using Llama 3.1, and GT API. Both translations are then passed to the *Evaluation Coordinator*. The *Evaluation Coordinator* uses GPT-4o to evaluate the LLM's initial translation and the GT translation based on the following criteria—*accuracy, fluency, terminology, and style*. Based on these metrics, the coordinator selects the most reliable initial translation for further refinement. This dual approach aims to leverage the strengths of both LLMs and GT, reducing the likelihood of initial translation errors, especially in handling complex linguistic structures.

Improving Initial Translation. The second layer improves the initial translation selected by the *Evaluation Coordinator*. The preferred translation is then passed to the *Proofreader* who provides a list of constructive feedback points to guide the improvement of the translation. This layer focuses on ensuring the translation matches the colloquial language of the target country and addresses key areas: *accuracy* (correcting errors such as additions, mistranslations, omissions, or untranslated text),

fluency (applying target language grammar, spelling, and punctuation rules, and eliminating unnecessary repetitions), *style* (reflecting the style of the source text and considering cultural context), and *terminology* (maintaining consistent use of terminology that reflects the source text domain) (Appendix B).

Also, within this layer, the *Editor* will use the *Proofreader*'s feedback to provide suggestions to improve the initial translation from the previous layer. This sublayer integrates the suggestions related to *accuracy, fluency, style,* and *terminology* to produce an improved translation.

Loss in Translation. Finally, the improved translation is passed to the *Editor-in-Chief* which uses GPT-4o to quantify how much the source text's meaning was lost in the translation process. To measure the loss in translation (LiT), first the *Editor-in-Chief* rates the translations on a scale of *Unacceptable*, *Satisfactory*, or *Excellent* for each of the following metrics—*accuracy (A), fluency (F), style (S), terminology (T),* and *cultural adaptation (CA)***.** *Unacceptable* was given a score of 1, *Satisfactory* a score of 2 and *Excellent* a score of 3. The scores are then used to calculate the loss in translation as follows:

$$LiT = 1 - \frac{(.4A + .2F + .2S + .1T + .1CA)}{3} \quad \textbf{(1)}$$

In translation, up to 10% loss is expected, so within the workflow if LiT is greater than 0.1, then the translation is passed back to the editor with the reasoning that the *Editor-in-Chief* provides (McKown et al., 2020). The translation process is further refined through an iterative loop, continuing until the desired quality is achieved or a maximum of $k = 3$ iterations is reached. The value of k was chosen to provide sufficient opportunities for improving the translation while balancing the cost of API usage and computational resources.

3.2 Tool Utilization

An essential component of the methodology is the utilization of external tools to enhance translation quality. By incorporating the GT API in the initial translation layer, the methodology benefits from an additional reference point. The *Evaluation Coordinator* compares the translations with each other, selecting the translation that best aligns with the source text. This approach mitigates initial translation errors and leverages existing translation technologies to bolster the overall effectiveness of the system.

3.3 Evaluation

The quality of the translations was assessed using both quantitative and qualitative measures. These metrics were applied to translations from GT, the Baseline model, and MATT, providing a comparison of MATT's performance against existing translation workflows.

BLEU score. The only quantitative measure used in this study is the SentencePiece BLEU (*sp*-BLEU) score, a variant of the BLEU metric that evaluates translations in the space of sentence pieces, as proposed by Goyal et al. (2022). Unlike traditional BLEU, which operates on word-level n-grams, *sp*-BLEU works on sub-word units. This provides an objective measure of the translation's closeness to human translations when comparing translations word to word. The *sp*-BLEU scores are limited to short text data, since ground truth human translations for the long text data were not available.

Bilingual Human Evaluation. One of the qualitative measures was Bilingual Human Evaluation (BHE), where the evaluators select the better response (Boubdir et.al., 2023). For each language, there were ten evaluators who were fluent in both English and the target language, Vietnamese, Hindi, or Malayalam. Each evaluator was given a random sample of 20 English source texts and 2 translations for each text. One of the translations was generated by MATT. The other translation was generated by either Baseline or GT. Evaluators were instructed to compare the two translations and vote for their preferred one. If both translations are of equal quality, they have the option to select "tie." Additionally, if neither translation meets the quality standards, they can choose "none" as their preference.

LLM Evaluation. The second measure was LLM evaluation. Similar to BHE, the LLM evaluator was randomly given the source text and 2 translations and voted on which result was preferred. The source texts given were a random sample of 100 short chunks and 10 long chunks. The BHE and LLM evaluations can result in one of four outcomes: a win for MATT, a win for the alternative translation, a tie (both are good), or neither (both are bad). In addition to choosing the preferred translation, the LLM gave a chain of thought reasoning for its decision. In this report the results from GPT-4o was used as the evaluator but appendix D compares the results with Claude and Gemma.

Between the three evaluation metrics, BHE is considered the gold standard for measuring translation quality and is the most heavily weighted when comparing MATT's performance to the Baseline model and the popular method, GT. Together, these three metrics provide a robust assessment of the final translation output, enabling a comprehensive comparison between the new workflow, the Baseline, and GT.

3.4 Integration of Components

By integrating these components, the multi-agent workflow layers, tool utilization, voting mechanism, evaluation process, and iterative improvement, MATT presents a potential for significant advancement in MT. It aims to offer a robust framework for producing translations that are accurate, fluent, stylistically appropriate, and with consistent terminology, particularly in low-resource language settings.

The focus on English to low-resource languages, including Vietnamese, Hindi, and Malayalam translation allows for testing the quality of the multi-agent translation workflow in a low-resource language context. These target languages were chosen due to the limited training data available for LLMs, which poses unique challenges and provides an opportunity to demonstrate the effectiveness of the proposed methodology.

4 The Data

This study employed 100 translation pairs randomly drawn from the Flores 101 benchmark to assess the translation capabilities of MATT on short texts and 10 long English texts from Open WebTexts2 and Research Paper datasets to assess its performance on long text chunks.

The Flores 101 benchmark is comprised of 3,001 sentences sourced from English Wikipedia, covering a variety of topics. On average, each sentence contains approximately 20 words. These sentences are derived from 1,175 distinct articles across three domains: "WikiNews," "WikiJunior," and "WikiVoyage" (Goyal et al., 2022). Some of the topics included in the dataset are crime, science, and politics. All 3,001 English sentences have been translated by professional human translators into 101 low-resource languages. However, this study specifically focuses on Vietnamese, Hindi, and Malayalam as the low-resource target languages. The sentences in the Flores data were first translated by a human translator and then sent to other human translators to edit the original translations. Then each translation was scored based on grammar, punctuation, spelling, capitalization, addition or omission of information, mistranslation, unnatural translation, untranslated text, and register (Goyal et al., 2022). If the translation passes the quality threshold of 90%, then the translation is kept, otherwise the process repeats until the threshold is met. The robust translation workflow of the dataset ensures that the target language translations serve as quality ground truths for evaluating the performance of the multi-agent workflow (Goyal et al., 2022).

The long text chunks included research papers and news articles covering topics like science, politics, and pop culture. Since these English texts lacked translation pairs in the target language, their evaluations were limited to the qualitative methods: BHE and LLM evaluation.

5 Results

Table 1. Comparison of *sp*-BLEU scores between different MT workflows and mechanisms.

	VIETNAMESE	HINDI	MALAYALAM
Baseline	0.2471	0.1390	0.0158
MATT	0.2952	0.1681	0.0231
GT	0.3886	0.2548	0.0728

Table 1 shows the *sp*-BLEU scores for the translations in the 3 target languages. It is evident that GT has the highest scores, regardless of language. For Vietnamese, GT has the highest *sp*-BLEU score of 0.3886 and for Malayalam it has the highest score of 0.0728. The difference in the highest scores between the two languages show that MT quality for Vietnamese is much better than for Malayalam. The MATT model outperforms the Baseline across all languages. However, the improvement is different for each language. For Vietnamese and Hindi, the improvement is 0.05 and 0.03 respectively, but for Malayalam, the improvement is less than 0.01.

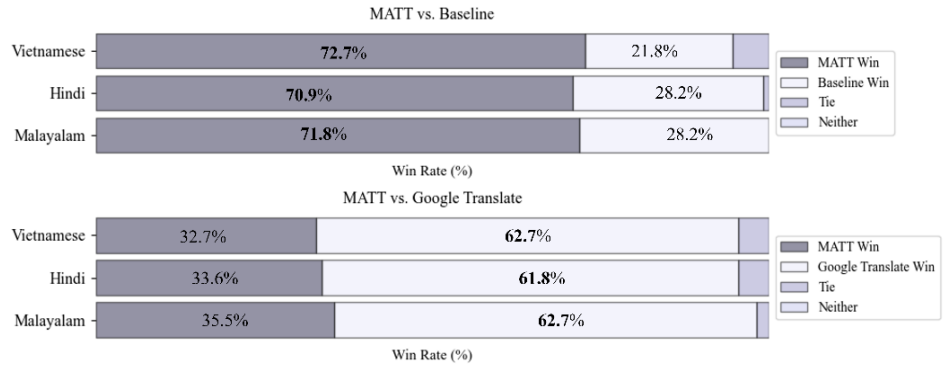

Fig. 2. Both plots show the results from the LLM Evaluation. The plot above compares the performance between MATT and Baseline (above). The plot below compares the performance between MATT and GT (below).

Both plots in Figure 2 highlight a consistent win rate for MATT across all three languages for their respective comparisons. In the top plot of Figure 2, which reflects the LLM evaluation results, MATT translations were preferred over the Baseline approximately 70% of the time across all three languages, indicating a strong preference for MATT's output. This significantly exceeds the preference rate observed in the BHE comparison between MATT and Baseline in the top plot of Figure 3.

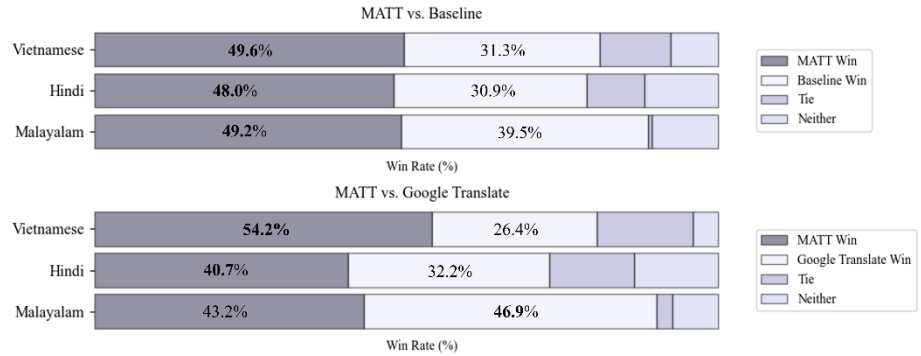

Fig. 3. Both plots show the results from the BHE. The plot above compares the performance between MATT and Baseline (above). The plot below compares the performance between MATT and GT (below).

The bottom plot in Figure 2, shows that the LLM evaluator preferred GT about 62% of the time across all three languages. On the other hand, MATT achieved a win rate of around 34% across all three languages. These results differ substantially from those of the BHE, where MATT had a much higher preference rate. The discrepancy between the two evaluation methods suggests differences in how LLMs and human evaluators perceive translation quality, emphasizing the importance of considering both automated and human perspectives in assessing performance.

The first plot in Figure 3 above shows the BHE performance for MATT vs. Baseline. For all three languages, it is shown that MATT was preferred by human evaluators. MATT was most strongly preferred for Vietnamese with an 18-point difference, and it was preferred for Malayalam with a 10-point difference. The bottom plot in Figure 3 shows the BHE performance for MATT vs GT. Similar to MATT vs Baseline, for Vietnamese and Hindi, MATT was preferred. For Vietnamese the difference is more significant with a 27-point difference, whereas for Hindi the preference was not as strong with a 6-point difference which is also less than the 11-point difference seen when comparing MATT and Baseline. For Malayalam, however, GT was preferred slightly with a 3-point difference. Further investigation must be carried out to understand the disjointed nature of the results from all the evaluation methods.

6 Discussion

6.1 Interpretation

From the BHE, the strong preference towards MATT across all three low-resource languages compared to Baseline indicates that the iterative and collaborative nature of the architecture can help improve LLMs performance on the complex translation task. Further, the noticeable favor towards MATT compared to GT for Vietnamese and Hindi indicates that LLMs may be able to outperform current translation systems for certain languages. However, the preference for GT over MATT for Malayalam shows that LLMs may not be able to outperform in certain low-resource languages because of the lack of available training data. Due to the inherent complexity of translation tasks, human evaluation remains the gold standard and is regarded as the most critical method in this research. However, there is an evident difference in the results obtained from the other two evaluation methods.

The increase in *sp*-BLEU scores for MATT compared to the Baseline across all languages agrees with the results from BHE. This cohesion between results demonstrates the quantitative improvements resulting from the adjustments and enhancements in MATT. Notably, Vietnamese, which exhibits the strongest preference for MATT, shows the largest improvement in *sp*-BLEU scores, while Malayalam, with the weakest preference for MATT, has the smallest improvement. This suggests that incorporating GT as a tool significantly enhances MATT's performance over Baseline. However, the varied improvements across languages indicate that translation quality is influenced and limited by the amount of training data available to the tools, LLMs and GT.

Although GT achieves higher *sp*-BLEU scores due to its reliance on word-to-word translations, BHE often prefers MATT over GT for Vietnamese and Hindi. This discrepancy arises because *sp*-BLEU scores, which measure n-gram overlap with ground truth translations, cannot account for the subjectivity and variability in translation. MATT and Baseline, which use LLMs to translate meaning rather than exact words, are less favored by the quantitative metric but align better with human

preferences, as seen in BHE results. For example, the phrase *"bed and breakfast"* is translated as *"dịch vụ giường ngủ và bữa sáng"* by GT, *"giường ngủ và bữa điểm tâm"* by MATT, and *"giường ngủ và bữa sáng"* in the ground truth from human translators. While the terms for *"breakfast"* (*"bữa điểm tâm"* and *"bữa sáng"*) are synonyms and convey the same meaning, GT achieves a higher *sp*-BLEU score (0.32) compared to MATT (0.19) because it uses the exact wording found in the human reference translation. Unlike Vietnamese and Hindi, Malayalam has a low *sp*-BLEU score and performs comparably in MATT and GT. This undistinguishable performance is likely a result of the poor performance of GT as a tool to translate Malayalam evidenced by GT's low *sp*-BLEU score.

The results from the BHE and LLM evaluations reveal important differences in how translation quality is perceived by human evaluators and LLM-based assessments. BHE results consistently show that MATT is preferred over the Baseline across all three languages, reflecting its ability to produce translations that better align with human expectations for fluency, style, and semantic accuracy. The LLM evaluator consistently favors MATT with a 70% preference across all languages. In contrast, BHE shows more variability in win rates across the classes ('MATT', 'Baseline', 'Tie', and 'Neither') depending on the language. The lack of variability indicates LLM's inadequacy in judging translations for these languages and potential opportunity for further prompt engineering.

LLM's difficulty in being a final evaluator is further seen in the strong preference that the LLM showed to GT over MATT. The LLM's results differ significantly from the BHE findings. While GT excels in producing literal translations that align closely with reference texts, MATT's approach focuses on capturing the intended meaning and delivering translations that resonate better with human evaluators.

Table 2. On the left is the translation generated by MATT, whereas translation on the right is generated by GT.

Source text: "You need to notice the victim's position as you approach him or her and any automatic red flags."	
MATT	GT
"Bạn cần chú ý đến vị trí của nạn nhân khi đến gần họ và mọi *dấu hiệu cảnh báo tức thời."*	"Bạn cần chú ý đến vị trí của nạn nhân khi bạn tiếp cận họ và bất kỳ *dấu hiệu cảnh báo tự động* nào."

For example, Table 2 shows that while both translations convey the main idea of the source text, there is a key difference in the interpretation of the term "automatic red flags." The LLM evaluator preferred the GT version, which translates "automatic red flags" to "dấu hiệu cảnh báo tự động," directly rendering "automatic" as "tự động," meaning something that operates automatically. However, this interpretation misses the intended nuance of the source text, where "automatic red flags" refers to immediate, spontaneous warning signs rather than something functioning automatically. In contrast, the MATT translation uses "dấu hiệu cảnh báo tức thời," which translates to "immediate warning signs," capturing the intended meaning of "automatic" as something one notices instinctively or spontaneously.

Therefore, the fact that the LLM evaluator preferred GT's translation raises concerns about its ability to capture contextual nuances in translation evaluation. The preference for a word-to-word translation, as seen in the GT version, suggests that the LLM evaluator prioritizes literal accuracy over semantic understanding. This limitation highlights the potential improvements in the prompt given to the LLM and in the LLM's ability to evaluate contextual meaning. These results suggest that while automated metrics can provide valuable insights, human evaluations remain critical for assessing the nuanced aspects of translation quality, particularly in low-resource languages.

While the findings from all three methods are not consistent, the findings from BHE and *sp*-BLEU scores suggest that MATT represents a meaningful improvement over both GT and the Baseline, particularly when augmented with the *Evaluation Coordinator*, GT, and *Editor-in-Chief*. Although the improvements are not overwhelming in every comparison, MATT consistently performs better and is preferred in most scenarios. This underscores the potential of MATT in enhancing translation quality, especially in more complex translation tasks where precision, clarity, and nuance are essential, such as in legal, medical, or technical research and documentation

6.2. Interesting Aspects

One of the most fascinating discoveries from the evaluation was how implementing GT as a tool and as an independent agent affects the performance of MATT. Although the LLM was assigned to be the main translator that could use GT as a reference tool, it would closely imitate the GT output and then attempt to improve upon it. As a result, translations that used GT as a reference often ended up being very similar to the original GT output, which can sound awkward and overly literal, almost like a word-to-word translation.

However, the implementation of GT as an independent translator alongside the LLM translator and the addition of the *Evaluation Coordinator* had a profound effect. Not only did it prevent the LLM from simply following the GT output, but it also selected a more appropriate initial translation. When the source text was more informal, the *Evaluation Coordinator* tended to choose the LLM-generated translation as the starting point. On the other hand, when the source text was more formal, such as in a peer-reviewed paper, the *Evaluation Coordinator* often selected the GT output as the initial translation. From there, the reflection agents acted as proofreader and editor, refining and improving the translation to ensure that the loss in meaning was kept below 10%. This dynamic process of selecting and refining translations demonstrates the versatility of the multi-agent system in handling a wide range of text complexities and contexts.

6.3. Limitations

The research faced several limitations due to the use of LLMs as a core component of the multi-agent workflow. A significant latency was encountered when translating the entire Flores dataset, leading to the decision to use a random sample of 100 sentences

instead of the full 3,001. While this reduction improved computational efficiency, it resulted in less robust findings due to the smaller dataset. Cost constraints also played a role in this decision. Most of the workflow utilized Llama 3.1, while GPT-4o was employed as the *Evaluation Coordinator* and *Editor-in-Chief*. However, using GPT-4o on a larger dataset would have exponentially increased the costs.

Another limitation was the inconsistency of the LLMs' outputs. When prompted to translate sentences or provide ratings, LLMs, like humans, did not always produce consistent results, and at times the quality was noticeably poorer. As a result, the quality of the response was limited by the researchers' prompt engineering capabilities (Appendix C). This variability led to small inconsistencies in the overall results. Additionally, the quality of the translations was constrained by the data on which the LLMs were trained. All three are low-resource languages, meaning the models had less training data compared to more widely spoken languages like Spanish. This lack of training data affected the effectiveness of the multi-agent translation workflow which was especially evident for Malayalam.

Evaluating the quality of translations and the degree of meaning lost in translation presented further challenges due to the absence of a standardized metric and the subjective nature of translations. The workflow employed a rubric assessing accuracy, fluency, style, terminology, and cultural context. *Sp*-BLEU scores, human and LLM evaluations were used for final assessments. However, the *sp*-BLEU scores skewed the results; for instance, while GT achieved the highest *sp*-BLEU scores, human evaluators generally rated its translations as inferior to those produced by MATT. Furthermore, the short and long texts were combined for all the evaluations, however, with the original Baseline development by Ng, there was a difference in performance for long and short texts. So, there may be a difference in performance of MATT vs the other translations for long vs short texts which was not identified. Additionally, the long texts were not able to be measured using *sp*-BLEU scores which may change the results.

Lastly, the research was constrained by the quality of human evaluators. While the individuals involved considered themselves fluent in both the target languages and English, colloquial fluency does not necessarily equate to proficiency in translation evaluation, introducing potential bias in the assessment process.

6.4 Ethics

The use of MATT in complex tasks, such as translation, raises significant ethical considerations. While translation is essential for many types of documents, particularly in legal, medical, and governmental contexts, the use of MATT powered by open-source LLMs introduces several ethical dilemmas when handling confidential or sensitive materials. Open-source LLMs often rely on data that is made publicly available or scraped from the internet, raising concerns about data privacy and confidentiality. When these models are used to translate sensitive documents, there is a risk that proprietary or confidential information may be exposed, either through the models' internal data processing or through third-party access during inference. Open-source LLMs may not always have robust security measures in place, potentially making them vulnerable to exploitation or leaks. This poses an ethical issue, as the inadvertent disclosure of sensitive information could have legal and reputational consequences for organizations and individuals alike.

Another ethical concern relates to data provenance and the possibility of unintentional bias. Translations are only as good as the LLMs and the LLMs are only as good as the data they were trained on, but the training data for these models is not always transparent. The lack of clarity around their sources could lead to translations that inadvertently reflect harmful biases, particularly in contexts involving marginalized communities. When using MATT for tasks like translation, ensuring that ethical safeguards are in place becomes crucial to prevent misuse and unintended harm.

6.5 Future Research

There are several areas of research that can be explored based on the findings of this study. One area could be to test multiple LLMs to gauge relative performance and latency using the same architecture. For low-resource languages, the amount of target language data the LLMs were trained on affects performance, so fine-tuning the LLM on specific target languages may have interesting results. Another is to assess the proposed architecture on other low and high-resource languages and further prompt engineering to create more consistent results. Additionally, exploring the integration of human-agent collaboration, where human expertise and machine efficiency are combined, presents a promising avenue for enhancing the capabilities and effectiveness of MATT. Finally, scaling the research and using other evaluation metrics instead of mainly human evaluation may be beneficial in providing a broader perspective.

7 Conclusion

This study aimed to enhance a current multi-agent workflow, to improve the quality of translations for three low-resource languages, Vietnamese, Hindi, and Malayalam. The workflow consists of 2 main parts, the first layer generates an initial translation, then the second layer iteratively improves it for final output. Across all languages when comparing MATT vs Baseline, the Human Evaluators had a stronger preference towards MATT. When comparing to the current popular commercial translation system, Google Translate, MATT was more strongly preferred by human evaluators for Vietnamese and Hindi with virtually no difference between Google Translate and MATT for Malayalam. The improvement does not apply only to these three languages but also suggests that a multi-agent implementation of LLM's can significantly improve the quality of translations for many languages, specifically low-resource languages.

Acknowledgments. We extend our deepest gratitude to our advisors, Joaquin Dominguez and Dr. Nibhrat Lohia, whose support, encouragement, and expertise were invaluable throughout this project. Their guidance was integral to the completion of our research, and we are truly grateful for their dedication. We would also like to recognize Andrew Ng and his team for developing the translation agent, which served as a foundation for our research and allowed us to design an innovative translation architecture. Additionally, we thank the faculty and staff at SMU who provided

resources and insights that contributed to this project. Finally, we are grateful to our colleagues and fellow students who offered feedback and support along the way.

References

1. Boubdir, M., Kim, E., Ermis, B., Fadaee, M., & Hooker, S. (2023). Which Prompts Make the Difference? Data Prioritization for Efficient Human LLM Evaluation. https://doi.org/10.48550/arxiv.2310.14424
2. Brief History of Translation: Everything You Need to Know. (n.d.). https://www.languagenetworkusa.com/resources/blog/brief-history-of-translation-everything-you-need-to-know
3. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. https://doi.org/10.48550/arxiv.2005.14165
4. Carl, J. (2023, January 17). A guide for cost-effective English to Vietnamese translations. Medium; Medium. https://medium.com/@johncarlacade1/a-guide-for-cost-effective-english-to-vietnamese-translations-7d6f37f29aa
5. Clinchant, S., Jung, K. W., & Nikoulina, V. (2019, September 27). On the use of BERT for Neural Machine Translation. ArXiv.org. https://arxiv.org/abs/1909.12744
6. Gillan, C. (2020, March 5). How many people speak English and where is it spoken? Lingoda. https://www.lingoda.com/blog/en/how-many-people-speak-english/#How-many-people-speak-English
7. Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., & Fan, A. (2022). The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. Transactions of the Association for Computational Linguistics, 10, 522–538. https://doi.org/10.1162/tacl_a_00474
8. Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., & Zhang, X. (2024, January 21). Large Language Model based Multi-Agents: A Survey of Progress and Challenges. ArXiv.org. https://doi.org/10.48550/arXiv.2402.01680
9. How Accurate is Google Translate? (n.d.). Slator. https://slator.com/resources/how-accurate-is-google-translate/#:~:text=Since%20its%20inception%20in%202006
10. Jiao, W., Huang, J., Wang, W., Wang, X., Shi, S., & Tu, Z. (2023a). ParroT: Translating During Chat Using Large Language Models. https://doi.org/10.48550/arxiv.2304.02426
11. Jiao, W., Wang, W., Jen-tse, H., Wang, X., Shi, S., & Tu, Z. (2023b). Is ChatGPT a Good Translator? Yes, With GPT-4 As the Engine. arXiv.Org. https://doi.org/10.48550/arxiv.2301.08745
12. Kocmi, T., & Federmann, C. (2023). Large Language Models Are State-of-the-Art Evaluators of Translation Quality. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2302.14520
13. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Abdelrahman, M., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv.Org. https://doi.org/10.48550/arxiv.1910.13461
14. Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Tu, Z., & Shi, S. (2023, May 30). Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. ArXiv.org. https://doi.org/10.48550/arXiv.2305.19118
15. Linguistics, 12, 229–246. https://doi.org/10.1162/tacl_a_00642
16. Lyu, C., Xu, J., & Wang, L. (2023, May 1). New Trends in Machine Translation using Large Language Models: Case Examples with ChatGPT. ArXiv.org. https://doi.org/10.48550/arXiv.2305.01181

17. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (2023, May 25). Self-Refine: Iterative Refinement with Self-Feedback. ArXiv.org. https://doi.org/10.48550/arXiv.2303.17651
18. Manasse, G. (2018, June 8). 4 Major Limitations of Human Translations. Bablic. https://www.bablic.com/blog/4-major-limitations-of-human-translations/
19. McKown, S., Acquadro, C., Anfray, C., Arnold, B., Eremenco, S., Giroudet, C., Martin, M., & Weiss, D. (2020). Good practices for the translation, cultural adaptation, and linguistic validation of clinician-reported outcome, observer-reported outcome, and performance outcome measures. Journal of Patient-Reported Outcomes, 4(1). https://doi.org/10.1186/s41687-020-00248-z
20. Mirela. (2024, January 4). The role of high-resource languages in NLP and localization. POEditor Blog. https://poeditor.com/blog/high-resource-languages/#:~:text=What%20are%20high%2Dresource%20languages
21. Mohamed, S. A., Elsayed, A. A., Hassan, Y. F., & Abdou, M. A. (2021). Neural machine translation: past, present, and future. Neural Computing and Applications, 33. https://doi.org/10.1007/s00521-021-06268-0
22. Ng, A. (2024a, March 20). Four AI Agent Strategies That Improve GPT-4 and GPT-3.5 Performance. Four AI Agent Strategies That Improve GPT-4 and GPT-3.5 Performance. https://www.deeplearning.ai/the-batch/how-agents-can-improve-llm-performance/?utm_campaign=The%20Batch&utm_source=hs_email&utm_medium=email\
23. Ng, A. (2024b, March 27). Multi-agent Design Patterns Part 2: Reflection. Multi-agent Design Patterns Part 2: Reflection. https://www.deeplearning.ai/the-batch/agentic-design-patterns-part-2-reflection/?ref=dl-staging-website.ghost.i
24. Oh, S., Lee, S. ah, & Jung, W. (2023). Data Augmentation for Neural Machine Translation using Generative Language Model. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2307.16833
25. Over 7000 languages are spoken in the world today, but not many are represented online - Consumers International. (2018, December 14). Www.consumersinternational.org. https://www.consumersinternational.org/news-resources/blog/posts/over-7000-languages-are-spoken-in-the-world-today-but-not-many-are-represented-online
26. Petrosyan, A. (2024, February 19). Most used languages online by share of websites 2024. Statista. https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/#:~:text=Common%20languages%20used%20for%20web%20content%202024%2C%20by%20share%20of%20websites&text=As%20of%20January%202024%2C%20English
27. Santy, S., Dandapat, S., Choudhury, M., & Bali, K. (2019, November 1). INMT: Interactive Neural Machine Translation Prediction (S. Padó & R. Huang, Eds.). ACLWeb; Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-3018
28. Schoening, S. (2023, September 19). Research vs Practice: How Accurate Is Google Translate? Phrase. https://phrase.com/blog/posts/is-google-translate-accurate/#:~:text=However%2C%20the%20accuracy%20levels%20ranged
29. Shinn, N., Labash, B., & Gopinath, A. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. ArXiv (Cornell University), 1(1). https://doi.org/10.48550/arxiv.2303.11366
30. Stahlberg, F. (2020). Neural Machine Translation: A Review. The Journal of Artificial Intelligence Research, 69, 343–418. https://doi.org/10.1613/jair.1.12007
31. The Demographics, Linguistics, and Important of Learning English in Vietnam (n.d). Lisa Helton https://americaneducationinternational.com/the-demographics-linguistics-and-importance-of-learning-english-in-vietnam/

32. Toshevska, M., & Gievska, S. (2021). A Review of Text Style Transfer Using Deep Learning. IEEE Transactions on Artificial Intelligence, 3(5), 669–684. https://doi.org/10.1109/TAI.2021.3115992
33. Varshney, T. (2023, November 30). Introduction to LLM Agents. NVIDIA Technical Blog. https://developer.nvidia.com/blog/introduction-to-llm-agents/
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. https://doi.org/10.48550/arxiv.1706.03762
35. Wang, J., Wang, J., Athwart, B., Zhang, C., & Zou, J. (2024). Mixture-of-Agents Enhances Large Language Model Capabilities. ArXiv.org. https://doi.org/10.48550/arXiv.2406.04692
36. Wang, W., Zheng, V., Yu, H., & Miao, C. (2019). A Survey of Zero-Shot Learning: Settings, Methods, and Applications. ACM Transactions on Intelligent Systems and Technology, 10(2), 1–37. https://doi.org/10.1145/3293318
37. Wang, Y., Sun, Z., Cheng, S., Zheng, W., & Wang, M. (2023, May 28). Controlling Styles in Neural Machine Translation with Activation Prompt. ArXiv.org. https://doi.org/10.48550/arXiv.2212.08909
38. Wang, Z., Pang, Y., & Lin, Y. (2023). Large Language Models Are Zero-Shot Text Classifiers. https://doi.org/10.48550/arxiv.2312.01044
39. Wu, M., Yuan, Y., Haffari, G., & Wang, L. (2024, May 20). (Perhaps) Beyond Human Translation: Harnessing Multi-Agent Collaboration for Translating Ultra-Long Literary Texts. ArXiv.org. https://doi.org/10.48550/arXiv.2405.11804
40. Yang, S., Wang, Y., & Chu, X. (2020). A Survey of Deep Learning Techniques for Neural Machine Translation. arXiv.Org. https://doi.org/10.48550/arxiv.2002.07526
41. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. ArXiv:1904.09675 [Cs]. https://arxiv.org/abs/1904.09675
42. Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., & Li, L. (2023, May 1). Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. ArXiv.org. https://doi.org/10.48550/arXiv.2304.04675

Appendix

A. Baseline Model Architecture

In this research, the Baseline model used for comparison with MATT is the T*ranslation_Agents* workflow developed by Andrew Ng and his team. Released in June 2024, this innovative approach to MT employs a reflection agentic workflow. The methodology utilizes LLMs to perform translations, reflect on their outputs and lastly, refine them based on this self-assessment. While no official flowchart has been released, its architecture can be conceptualized as follows:

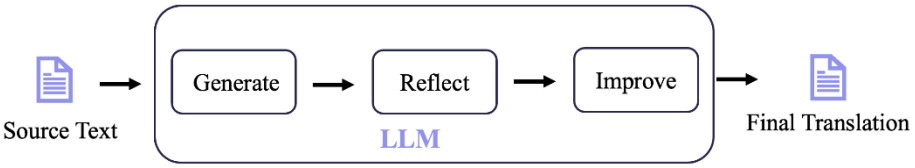

Fig. 4. This flowchart highlights the main components in the Baseline model

It begins with a LLM translating a text from the source language to the target language. The translated output is then subjected to a reflection phase, where the LLM evaluates and generates suggestions for its own translation. It identifies potential areas for improvement, such as inaccuracies, inconsistencies, or stylistic adjustments and generates constructive suggestions for refining the translation. Finally, the model incorporates these suggestions to produce an improved and more polished translation.

The additional layer of self-critique makes it more effective than a simple LLM-generated translation. However, despite its iterative architecture, the model is still susceptible to issues such as mistranslations and hallucinations, where the LLM generates content that is factually inaccurate or unrelated to the source text. These hallucinations can propagate through the reflection stage and subsequently influence the refined output, as the process relies entirely on the LLM's internal capabilities to both critique and correct itself.

B. Metrics

Improving Translation Metrics. In both frameworks, the Baseline and MATT, the initial translation is reviewed and improved based on the following metrics:

Table 3. These metrics are implemented in the prompts for the LLM to critique the initial translations and provide suggestions for improvement.

Metrics	
(i)	Accuracy (by correcting errors of addition, mistranslation, omission, or untranslated text),
(ii)	Fluency (by apply {target_lang} grammar, spelling, punctuation rules, and ensuring there are no unnecessary repetitions),
(iii)	Style (by ensuring the translations reflect the style of the source text and takes into account any cultural context),
(iv)	Terminology (by ensuring terminology use is consistent and reflects the source text domain; and by only ensuring you use equivalent idioms {target_lang})

These metrics are incorporated as part of the reflection agent in the Baseline model and are implemented in the *Proofreader* agent in MATT. The primary objective of these agents in the Baseline and MATT is to facilitate a comprehensive review of the initial translation, identifying areas for improvement and applying enhancements to elevate the overall quality of the translation.

Loss in Translation Evaluation Metrics. In addition to the four metrics from the previous layer, the *Editor-in-Chief* evaluates the loss in cultural adaptability within the improved translations. This metric is introduced to address the limitation in handling cultural nuances in LLMs and MT in general.

Tables 4 to 8 present the three-level rating scale and corresponding descriptions used by the *Editor-in Chief* to evaluate the loss for each metric.

Table 4: Rating Scale and Descriptions for Evaluating Accuracy

Rating	Description
Excellent	Completely accurate with no deviations from the original meaning; all key details and nuances are conveyed exactly as intended.
Satisfactory	Largely accurate with only minor errors that do not affect overall comprehension.
Unacceptable	Significant distortions of meaning; important information is incorrect or missing; mistranslations leading to misunderstanding of the content.

Table 5: Rating Scale and Descriptions for Evaluating Fluency

Rating	Description
Excellent	Completely fluent and natural; no grammatical errors or awkward expressions; flows as if originally written in the target language.

Satisfactory	Mostly fluent with only a few unnatural expressions or isolated grammatical errors; reads smoothly overall.
Unacceptable	Extremely awkward and unnatural; numerous grammatical mistakes; difficult to read and follow.

Table 6: Rating Scale and Descriptions for Evaluating Style

Rating	Description
Excellent	The original style, tone, and voice are fully maintained; the translation mirrors the source text perfectly in its stylistic delivery.
Satisfactory	The style is well preserved with only slight deviations; the translation maintains the appropriate tone and reflects the original text's voice effectively.
Unacceptable	Significant loss of the original style; tone and voice are inappropriate or completely inconsistent with the source text.

Table 7: Rating Scale and Descriptions for Evaluating Terminology

Rating	Description
Excellent	Terminology is consistently accurate and precise throughout; all domain-specific terms are used correctly and appropriately.
Satisfactory	Terminology is accurate and consistent, with only minor and isolated errors that don't impact the overall meaning or coherence.
Unacceptable	Incorrect or inconsistent use of key terminology; domain-specific terms are either mistranslated or completely missing.

Table 8: Rating Scale and Descriptions for Evaluating Cultural Adaptability

Rating	Description
Excellent	Cultural references, idioms, and context are perfectly adapted to the target language.
Satisfactory	Minor cultural nuances are missed but do not affect overall comprehension.
Unacceptable	Cultural elements are entirely mishandled, causing significant loss of meaning.

C. LLM Evaluation Prompts

Different LLMs were used to evaluate MATT's performance, with GPT-4o showing results closest to human evaluation when comparing MATT and Baseline. The prompt is demonstrated as follows:

```
“““
You are an expert in evaluating translation quality. Your task is to carefully read a source text and two translations from {source_lang} to {target_lang}, and choose which translation is best keeps the meaning of the {source_text}.
The source text and two translations, delimited by XML tags <SOURCE_TEXT></SOURCE_TEXT>, <TRANSLATION_A></TRANSLATION_A>, and <TRANSLATION_A></TRANSLATION_A> are as follows:

<SOURCE_TEXT>
{source_text}
</SOURCE_TEXT>

<TRANSLATION_A>
{shuffled_translation_a}
</TRANSLATION_A>

<TRANSLATION_B>
{shuffled_translation_b}
</TRANSLATION_B>

When choosing between Translation A and Translation B, pay attention to which translation keeps the meaning of the {source_text} considering the following:
(i) accuracy (no deviations from the original meaning; all key details and nuances are conveyed exactly as intended),
(ii) fluency (fluent and natural; no grammatical errors or awkward expressions; flows as if originally written in the {target_lang}),
(iii) style (maintained the original style, tone, and voice of the {source_text}; the translation mirrors the source text perfectly in its stylistic delivery),
(iv) terminology (terminology use is consistent and reflects the source text domain; and equivalent idioms in {target_lang} are used)
(iv) cultural adaptability (Cultural references, idioms, and context are perfectly adapted to the {target_lang} and {country}.
Write which translation is preferred or if it's a tie or none and explain why.
”””
```

D. Comparison of Different LLM Evaluations

For the evaluation of MATT's performance, additional LLMs were utilized alongside OpenAI, including Gemma and Claude. Tables 9-11 highlight the differences in how these LLMs assessed MATT's translations compared to the baseline and GT translations in three target languages.

Table 9. The comparison of the LLM Evaluation in Malayalam

LLMs	Gemma	Claude	OpenAI
MATT wins over GT	55.56	48.54	35.45
MATT wins over Baseline	35.19	70.91	71.82

According to Table 9, when assessing MATT against GT, Gemma favored MATT in 55.56% of cases, followed by Claude (48.5%), and OpenAI (35.45%). Conversely, when comparing MATT to the Baseline, OpenAI showed the strongest preference for MATT at 71.82%, while Claude and Gemma recorded 35.45% and 35.19%, respectively. Ironically, the fact that Gemma chose MATT over GT in Malayalam contradicts both the *sp*-BLEU scores and BHE results. In these assessments, MATT performed best in Vietnamese and worst in Malayalam, making Gemma's preference for MATT in Malayalam outlier when compared to the broader evaluation metrics.

Table 10. The comparison of the LLM Evaluation in Hindi

LLMs	Gemma	Claude	OpenAI
MATT wins over GT	38.89	40.00	33.64
MATT wins over Baseline	50.46	61.37	70.91

Among the LLMs, Claude consistently underestimated MATT's performance against GT. This evaluation does not align with either BHE or the *sp*-BLEU scores, which indicate that MATT performs as well as, if not better than, the competitions across all languages. Notably, both BHE and *sp*-BLEU scores suggested that MATT performs better in Hindi than in Malayalam. However, Claude's assessment showed otherwise, as it rates MATT's performance in Hindi lower than in Malayalam. This misalignment highlights significant discrepancies between Claude's evaluations and other evaluation metrics, further emphasizing the limitations of relying on this LLM for translation quality assessment.

Table 11. The comparison of the LLM Evaluation in Vietnamese

LLMs	Gemma	Claude	OpenAI

MATT wins over GT	48.62	32.73	32.73
MATT wins over Baseline	63.31	72.22	72.72

The discrepancies among different LLM evaluation tools raise concerns about their reliability and consistency. Although OpenAI demonstrated greater alignment with human judgments, particularly in preferring MATT over the Baseline, the stronger preference for GT translations by most LLMs suggests that these tools may lack nuanced interpretability in certain linguistic contexts. These findings highlight the need for further investigation into LLM evaluation processes, including improvements in prompt design, to develop less biased and more reliable tools that better reflect human evaluations. Such advancements would enhance the utility of LLMs as evaluation tools and improve their consistency across various languages and contexts.

E. LLM Models

In this study, "Llama-3.1-70b-versatile" model was used for the translator, proofreader, and editor agents. Also, GPT-4o was used for the evaluation coordinator and *Editor-in-Chief* agents.