MAML: Multi-Agentic and Multi-Level CoT for LLM-Based Automatic Subtitle Translation Evaluation

Anonymous ACL submission

Abstract

Subtitle translation is crucial in ensuring global accessibility, particularly for creative content such as films and television. However, the manual translation is labor-intensive and time-consuming, often requiring linguistic and cultural adaptation. While post-editing accelerates the translation process, effective automatic evaluation methods are crucial to ensure fair and reliable quality assessment while minimizing human effort. We propose TEAM (Translation Evaluation and Assessment with Multi-agents), a novel agent-based evaluation metric designed to identify the most creatively aligned translations while preserving linguistic quality. TEAM assesses key factors such as cultural relevance, emotional tone, humor, and engagement, helping post-editors select and refine the best machine-generated translations. Additionally, we propose MLCoT (Multi-Level Chain-of-Thought), a simpler metric where multiple agents evaluate adequacy, fluency, and creativity. Experiments on English-Hindi and English-Spanish subtitles show that both TEAM and MLCoT outperform CometKiwi in preference ranking and correlation with humans.

1 Introduction

Subtitle translation plays a crucial role in making multimedia content accessible across languages and cultures (Pettit, 2009). While manual translation ensures high quality, it is costly, time-consuming, and labor-intensive process. Post-editing has emerged as a practical solution, allowing translators to refine machine-generated outputs instead of translating from scratch (Matusov et al., 2019; C. M. de Sousa et al., 2011). Studies have shown that post-editing significantly improves efficiency, reducing translation time while maintaining high quality. However, for post-editing to be effective, it requires an effective evaluation framework that ranks machine-generated translations based on quality. Figure 1 illustrates that human subtitle generators benefit from a pre-sorted list of options, reducing cognitive load and improving productivity. The ranking process relies on automatic evaluation metrics that assess adequacy, fluency, and contextual alignment.

No.	Subtitles & Translations	Score
1	**en:** How are you? Im great! **es:** Cómo estás? *(How are you?)*	0.84
2	**en:** We, uh...we caught a break in the Spiderwoman case. **es:** Encontramos un avance... en el caso de Spiderwoman. *(We found a breakthrough...* *in the case of Spiderwoman.)*	0.68
3	**en:** About to roll without you. **hi:** तुम्हारे बिना जाने वाला था। *(I was about to go without you.)*	0.6
4	**en:** Fifty Thousand! **hi:** पाँच लाख! *(Five lakhs!)*	0.86
5	**en:** Hey, yo, H should be expecting us. **hi:** अरे, यो, एच हमें इंतजार करना चाहिए। *(Hey, yo, H we should wait.)*	0.71

Table 1: Examples Illustrating the Limitations of the CometKiwi Metric.

Commonly used metrics like BLEU, ChrF, and COMET, while effective in many translation tasks, they rely on reference-based evaluations (Papineni et al., 2002; Popović, 2015, 2017; Mukherjee et al., 2020; Mukherjee and Shrivastava, 2023; Zhang* et al., 2020; Rei et al., 2020). However, these metrics are often unsuitable for real-time evaluations of subtitle translations, where true (*gold*) references are not always available and creative adaptations are needed. In contrast, metrics like CometKiwi (Rei et al., 2022) offer reference-free assessments and have been widely used

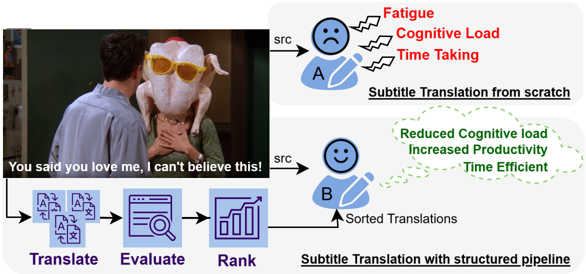


Figure 1: Subtitle translation task with and without structured pipeline. The sorted list simplifies the process for *Subtitle Generator B*, saving time and reducing cognitive load. The scene is source from MELD Dataset (Poria et al., 2019).

in reference-less translation quality estimation (QE) tasks (Blain et al., 2023; Zerva et al., 2024). Though CometKiwi is the SOTA QE metric, it does not perform well on informal, culturally rich, and creative contexts. Trained on formal news datasets, it often misjudges nuanced translations. Our experiments indicate that CometKiwi often assigns high scores to formal translations that are factually inaccurate and/or grammatically incorrect and low scores to accurate but creative subtitles. For instance, in Table 1, CometKiwi assigns a high score to incomplete (1), hallucinated (4), and incorrect translations (5) while penalizing correct subtitle translations (2 & 3). **These shortcomings underscore the need for a reliable evaluation method for the creative domain.**

To address these challenges, we introduce MLCoT (**M**ulti-**L**evel **C**hain-**o**f-**T**hought) and TEAM (**T**ranslation **E**valuation and **A**ssessment with **M**ulti-agents), designed to assess subtitle translations more effectively by leveraging large language models (LLMs) to evaluate key aspects such as cultural relevance, emotion, tone, humor, and engagement. Instead of relying on a single evaluation pass, MLCoT and TEAM evaluate quality at various levels of expertise, before aggregating their judgments. These structured approaches ensure a more reliable and context-aware ranking of translations, guiding post-editors toward the most suitable options. We explored GPT-4o-mini (Xie and Wu, 2024) and other publicly available models like Aya23 (Aryabumi et al., 2024), Gemma (Gemma, 2024) and Llama3.1 (LLama, 2024) to evaluate English-Hindi and English-Spanish movie subtitles (sec 4.1). Our experiments conclude that these methods outperform CometKiwi in terms of preference ranking (sec 5.1) and correlation with human assessments (sec 5.2).

Beyond entertainment, TEAMs framework can be adapted to other domains by defining the agent roles. By bridging the gap between linguistic accuracy and domain requirements, TEAM provides reliable automated evaluation, ultimately improving the quality of machine-generated subtitles. The key contributions of our work are summarized as follows:

1. **Translation Evaluation and Assessment with Multi-agents:** To the best of our knowledge, we are the first to introduce a novel multi-agent evaluation method for translation assessment.

2. **Improved Subtitle Generation Pipeline:** We propose a pipeline that includes preferred translations, further improving the efficiency of subtitle generation/post-editing.

3. **Leveraging LLMs for Context and Reasoning:** We explore several methods for utilizing the broader context and reasoning capabilities of large language models (LLMs) to assess subtitle quality.

2 Related Work

Subtitle translation has been extensively studied in both traditional and computational linguistics. Early research highlighted the linguistic and cultural challenges of subtitling, emphasizing accuracy, fluency, and adherence to space-time constraints (Gottlieb, 1997). While human translators ensure quality, this process is time-intensive, leading to increased interest in automation through machine translation (MT). Advancements in statistical (Koehn et al., 2003; Koehn, 2009) and neural MT (Vaswani, 2017) have improved subtitle translation efficiency. However, MT systems exhibit trade-offs (Koehn and Knowles, 2017; Läubli et al., 2018), some models excel in fluency but compromise fidelity, others maintain accuracy but sound overly formal, and LLMs often generate creative yet sometimes inconsistent translations (Court and Elsner, 2024). Despite ongoing research on improving MT's handling of idioms (Liu et al., 2023; Donthi et al., 2025) and cultural appropriateness (Adilazuarda et al., 2024), **achieving consistency in preserving these creative nuances in the subtitle, remains a persistent challenge** (Pedersen, 2005; Arenas and Toral, 2022).

To improve subtitle translation efficiency, post-editing has been explored as a viable approach. Matusov et al. reported a 37% increase in productivity for English-Spanish, while C. M. de Sousa et al. found that post-editing was 40% faster than human translation of English-Portuguese movie subtitles. While these findings suggest that post-editing boosts efficiency, it still depends on human expertise to refine MT outputs. **An automated evaluation and ranking system can address this challenge by pre-selecting high-quality translation options, reducing human effort while ensuring accuracy and cultural relevance.**

Evaluation metrics play a crucial role in assessing and ranking subtitle translations. Traditional metrics like BLEU (Papineni et al., 2002), ChrF (Popović, 2015, 2017), BERTScore(Zhang* et al., 2020), COMET(Rei et al., 2020) rely on reference translations, **making them impractical for evaluating subtitles in real-time**. Whereas, CometKiwi (Rei et al., 2022) is a reference-free metric that has been effectively used as a baseline in the recent WMT Quality Estimation (QE) Tasks(Blain et al., 2023; Zerva et al., 2024). However, as a supervised metric trained primarily on formal data, **CometKiwi struggles with informal text and cultural nuances.** These limitations highlight the need for novel evaluation metrics that incorporate cultural and contextual relevance along with adequacy and fluency.

With the rise of LLM-based evaluation, metrics like GEMBA (Kocmi and Federmann, 2023) show strong correlation when references are available, Lu et al. enhanced segment-level evaluation via prompt engineering using GPT-3.5-Turbo, Sato et al. fine-tuned GPT-4o-mini for WMT24 QE Task, achieving 1st place in English-Gujarati and English-Hindi, and 4th in English-Tamil and English-Telugu. These studies highlight the potential of LLM-based metrics as effective alternatives to humans.

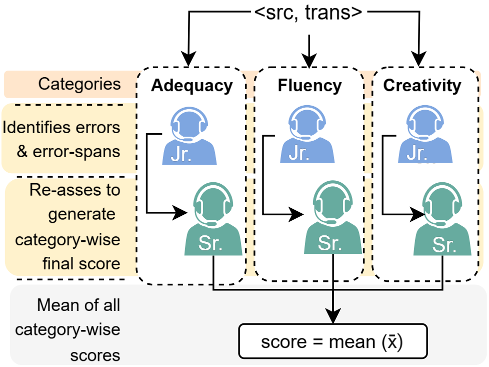

<src, trans>
Categories
Adequacy
Fluency
Creativity
Identifies errors & error-spans
Jr.
Re-asses to generate category-wise final score
Sr.
Mean of all category-wise scores
score = mean (x̄)

Figure 2: MLCoT prompting and score aggregation. Jr.: Junior Reviewer; Sr.: Senior Reviewer

Despite their strengths, LLMs struggle with structured reasoning when using CoT prompts, especially in smaller models (Wei et al., 2022). Multi-agent LLM frameworks (Wu et al., 2024) address this by distributing tasks among specialized agents. Building on this, we propose multi-agent based evaluation to assess subtitle translations. This structured approach ensures more reliable assessments, particularly informal and cultural nuances, overcoming the limitations of existing reference-free metrics like CometKiwi.

3 Our Approach

We present our approaches for leveraging LLMs to assess creative translations: Multi-Level CoT (MLCoT) and TEAM (Translation Evaluation and Assessment with Multi-agents).

3.1 CoT to MLCoT

Our work uses GEMBA (Kocmi and Federmann, 2023), an LLM-based SOTA metric, as a baseline. We improve its prompt (refer Appendix A Table 7) to address entertainment-specific challenges like tone, cultural context, and idiomatic expressions. Using this CoT prompt, the model first identifies and classifies errors (e.g., meaning, grammar, terminology, idiomatic expressions), assesses their severity, and categorizes them as major or minor. Finally, the model assigns an overall evaluation score, ranging from 0 to 100, reflecting the translation's quality.

We extend CoT to Multi-level CoT prompting to evaluate machine-generated subtitle translations at multiple levels (*junior-level and* *senior-level*) across three key categories: **1) Adequacy or Meaning Transfer, 2) Fluency, and 3) Creativity**. Figure 2 depicts the evaluation process, where initially, the Junior Reviewers independently assess the subtitle translations in terms of their respective category. Their primary role is to identify error spans, and further sub-categorize errors from a predefined error-list available for all three categories (refer Appendix A Table 8 & 9). The error spans, and errors identified by Junior Reviewers are reassessed by Senior Reviewers independently for the corresponding category, ensuring that the translations adhere to high linguistic and cultural standards appropriate for subtitles in entertainment content. The final evaluation score is derived by averaging the scores assigned by the Senior Reviewers. This multi-level evaluation approach ensures a comprehensive assessment, combining the initial insights of Junior Reviewers with the refined expertise of Senior Reviewers, ultimately improving the accuracy and quality of machine-generated subtitle translations.

3.2 TEAM: Translation Evaluation and Assessment with Multi-agents

After exploring CoT and MLCoT, we take a further step toward mimicking human judgment by introducing the TEAM framework, as shown in figure 3. Inspired by TransAgents (Wu et al., 2024), which employs a multi-agent framework for literary translations, we extend this to evaluate subtitle translations. We incorporate multiple specialized agents that leverage the human-like reasoning capability of LLMs (refer Appendix A.6) with automated processes to ultimately offer a thorough evaluation of subtitle translations, ensuring linguistic accuracy and cultural relevance, particularly for entertainment content such as movies and TV shows. The workflow of our architecture is detailed in Algorithm 1.

3.2.1 Agents

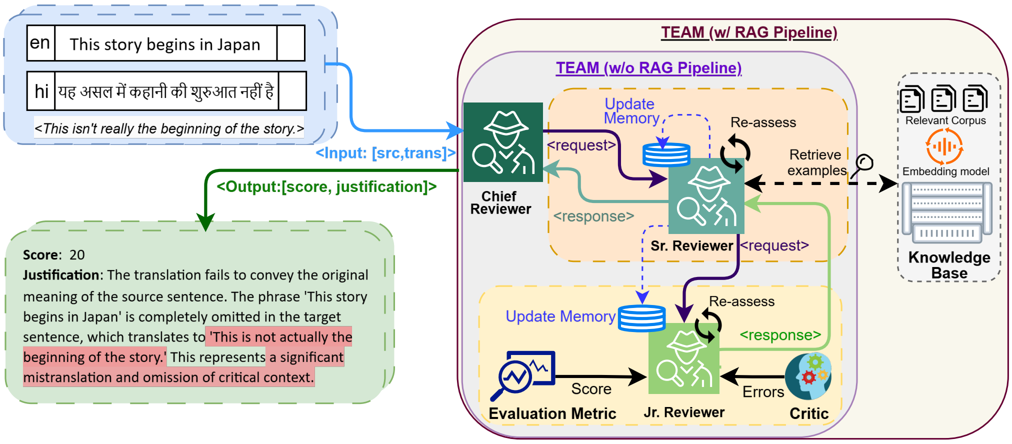
TEAM consists of a Chief Reviewer and three specialized agents: the Senior Reviewer, the Junior Reviewer, and the Critic focusing on entertainment-specific issues such as idiomatic expressions, humor, and cultural nuances, evaluating the translation's intended tone and context. The agent roles are described as follows:

- **Chief Reviewer** a)receives the source and translation, b)coordinates the flow between Senior Reviewer and Junior Reviewer, and c)outputs the score and justification.
- **Senior Reviewer** (Sr.Reviewer) provides an initial score and later, reassesses by considering feedback from the Jr.Reviewer (including errors identified by the Critic). Sr.Reviewer has access to relevant examples (sec 3.2.2) from the entertainment domain, such as movie and TV show subtitles, ensuring that the evaluation aligns with the domain requirements.
- **Junior Reviewer** (Jr.Reviewer) evaluates the translation and refines iteratively, considering the automatic quality estimation score and the critical errors identified by the Critic agent. This iterative process ensures a better comprehensive assessment of the translation.
- **Critic**: The Critic's role is to analyze the original dialogue and its translation. Their primary task is to identify errors (if any), including issues with accuracy, tone, cultural appropriateness, etc, compared to the original dialogue.

3.2.2 Extrinsic Knowledge Integration (RAG)

To enhance the capabilities of the Sr.Reviewer and make it function more like a manual translation evaluator, we integrate a vector database containing tuples of the form $u \equiv \langle$ source language, sentence, target language translation, errors, the severity of the error, the score assigned $\rangle$. This vector database is utilized in two stages:

- **Initial Retrieval:** Given a source sentence and its translation, denoted as $\langle x, y \rangle$, the Sr.Reviewer retrieves k-many relevant tuples $(u_1, u_2, \ldots, u_k)$ from the vector database. This provides an initial assessment of translation quality based solely on the semantics of $\langle x, y \rangle$.
- **Refined Evaluation:** After receiving the Jr.Reviewers assessment, which includes fine-grained error categories e_{jr},



Score: 20
Justification: The translation fails to convey the original meaning of the source sentence. The phrase 'This story begins in Japan' is completely omitted in the target sentence, which translates to 'This is not actually the beginning of the story.' This represents a significant mistranslation and omission of critical context.

Figure 3: Illustration of TEAM workflow

lp	en-hi		en-es	
movie	Creed	Pushpa	Creed	Spiderman
HC	100	100	100	100
MC	100	100	100	100
MF	100	100	100	100
Total	300	300	300	300

Table 2: Testset Statistics: A total of 1200 subtitle pairs across 4 movies and 3 translation types.

the Sr.Reviewer performs a second retrieval of k tuples. This retrieval considers both $\langle x, y\rangle$ and e_{jr} , offering a reliable external knowledge source to guide the final evaluation and score assignment.

4 Experimental set up

We assess machine-generated and human-generated movie subtitles using MLCoT (Sec:3.1), TEAM (Sec:3.2) and baselines metrics (CometKiwi and CoT(Sec:3.1)). **The main aim of our experiment is to identify the best-suited metric to use in the subtitle post-editing pipeline (Figure 1), i.e., which metric favors 'accurate-creative' subtitles over 'formal-literal' ones.**

4.1 Test Dataset

We use subtitles from popular movies, Pushpa and Creed for English-Hindi (en-hi), Creed and Spider-Man for English-Spanish (en-es), sourced from Opensubtitles[1], containing original English subtitles with human-written creative translations (HC) in Hindi and Spanish. To enhance the test corpus, we include **a)** ***machine-generated formal translations*** **(MF)**, using IndicTrans2 (Gala et al., 2023) for Hindi and NLLB (Team et al., 2022) for Spanish, providing accurate but literal translations[2] and **b)** ***machine-generated creative translations*** **(MC)**, by prompting LLMs (refer Appendix A Table 13) to generate translations that strike a balance between accuracy and creative expression, similar to humans. Table 2 reports our test data distribution.

4.2 Evaluation Approaches

We evaluated 1200 source-translation subtitle pairs by **GPT-4o mini** using *CoT*, *MLCoT*, *TEAM* and *TEAM w/o RAG*[3]. Given that GPT-4o mini is a commercial model, we later extended our experiments by utilizing publicly available, smaller models such as Gemma[4], Aya23[5], and Llama 3.1[6]. (Model details are reported in Appendix A Table 10). However, we observed that these models struggled with CoT prompts and failed to effectively identify errors, categorize, assign severity, and provide a final score. Interestingly, most of these models assigned identical scores to the majority of the samples, indicating their inability to evaluate accurately. Hence, we decided to use the proposed TEAM approach for our experiments involving the smaller models.

[1] https://www.opensubtitles.org/

[2] IndicTrans2 and NLLB are trained on large parallel corpora to generate formal, standard translations, focusing on precision rather than cultural adaptation.

[3] same as TEAM but without RAG pipeline

[4] https://huggingface.co/google/gemma-2-9b-it

[5] https://huggingface.co/CohereForAI/aya-23-8B

[6] https://huggingface.co/meta-llama/Llama-3.1-8B

Algorithm 1: Translation Assessment and Evaluation with Multi-Agents

```
Input: Source $src$; Translation $trans$
Output: Final evaluation $R$
$H \leftarrow [src, trans]$ $R \leftarrow \emptyset$ // Initialize the final evaluation
$m \leftarrow 0$ // Current round of evaluation
// Step 1: Initial Evaluation by Senior Reviewer
// Initial Retrieval of relevant examples using the RAG pipeline
$k\_examples \leftarrow KnowledgeBase(src)$
$sr\_score, sr\_justification \leftarrow sr\_reviewer(H, k\_examples)$
// Step 2: Initial Evaluation by Junior Reviewer
$jr\_initial\_score, jr\_justification \leftarrow jr\_reviewer(H)$
// Step 2.1: Retrieve translation errors from Critic
$critical\_errors \leftarrow crtic(H)$
// Step 2.2: Get a quality score by an automatic metric
$qe\_score \leftarrow EvaluationMetric(H)$
// Step 2.3: Detailed Evaluation by Junior Reviewer
$jr\_score, jr\_justification \leftarrow$
  $jr\_reviewer(H, jr\_initial\_score, qe\_score, critical\_errors)$
while $m \leq M$ do
    $m \leftarrow m + 1$
    // Update history
    $jr\_history \leftarrow jr\_score, jr\_justification, qe\_score, critical\_errors$
    // Step 3: Re-prompt Junior Reviewer with previous judgement
    $jr\_score, jr\_justification \leftarrow jr\_reviewer(H, jr\_history)$
    if $|prev\_jr\_score - jr\_score| \leq acceptable\Delta$ then
        Break // Stop iterating as the score difference is now acceptable
    // Step 3.2: Junior Reviewer iteratively adjusts score
    $jr\_score, jr\_justification \leftarrow$
      $jr\_reviewer(H, jr\_initial\_score, qe\_score, critical\_errors)$
// Step 4: Junior Reviewer sends feedback to Senior Reviewer
$H \leftarrow H + [jr\_score, critical\_errors, jr\_justification]$
// Step 5: Senior Reviewer re-assesses based on Junior Reviewer's feedback
// Refined Retrieval of relevant examples using the RAG pipeline
$k\_examples \leftarrow KnowledgeBase(src, critical\_errors)$
$sr\_score, sr\_justification \leftarrow sr\_reviewer\_assess(H, k\_examples)$
// Step 6: Senior Reviewer sends final evaluation to Chief Reviewer
$R \leftarrow sr\_reviewer\_finalize(sr\_score, sr\_justification)$
return $R$ // Return the final evaluation report
```

In addition, we evaluated using the SOTA reference-free metric, CometKiwi and further included a few reference-based metrics, BLEU, ChrF, MEE4, BERTScore and COMET; with human-generated subtitles (HC) serving as the reference (metric implementation details are mentioned in Appendix A Table 12).

5 Meta-Evaluation

We now assess how different metrics prioritize creative over formal subtitles through a comparative analysis of preference rankings, rank correlation coefficients, and pairwise accuracies.

5.1 Preference Ranking

We compute preference ranking percentages to examine the metric preferences among the three translation styles (HC, MC, MF), as shown in Table 3. For example, HC>=MC represents the percentage of instances where human-written creative translations (HC) are

			en-hi			en-es		
Metric	ref.	sup.	HC >= MC ↑	HC >= MF ↑	MC >= MF ↑	HC >= MC ↑	HC >= MF ↑	MC >= MF ↑
BLEU	✓	-	-	-	57	-	-	45
ChrF++	✓	-	-	-	47	-	-	44
BERTScore	✓	-	-	-	47.5	-	-	42
MEE4	✓	-	-	-	40.5	-	-	43.5
COMET	✓	✓	-	-	50	-	-	43.5
CometKiwi*	-	✓	44	41	45	51	43.5	59
CoT_GPT*	-	-	**52**	47	57	**59**	**67.5**	71
MLCoT_GPT	-	-	45.5	50.5	**62.5**	46	60	**71.5**
TEAM_GPT **(w/o RAG)**	-	-	44.5	50	61	48	57.5	70
TEAM_GPT	-	-	46.5	50.5	62	51	60	71
TEAM_Gemma	-	-	41.5	41.5	60.5	51	**65**	**75.5**
TEAM_Aya23	-	-	50.5	**55.5**	**65**	**59**	60.5	64.5
TEAM_Llama3.1	-	-	**62.5**	**65**	**65**	32	32.5	43.5

Table 3: Preference Ranking in percentages (%). Column-wise top % are underlined and highlighted. Baselines are marked with asterisk(*). ref.: Reference-based Metric; sup.: Supervised Metric.
(Eg: For en-es testset, 71% of times TEAM_GPT and 59% of times CometKiwi, have assinged higher or equal rank to Machine Generated Creative Translations (MC), in comparison with Machine Generated Formal Translations (MF).

ranked better than or equal to machine-generated creative translations (MC).

5.2 Comparing with Human Ranks

We conducted a human evaluation on randomly selected 75 en-hi source-translation pairs, where three native speakers ranked the subtitles (HC, MF, and MC) from 1 (best) to 3 (worst) (refer Appendix A.1). Table 4 reports the Intra-Class Correlation (ICC), depicting the agreement among the evaluators (Shrout and Fleiss, 1979). The 0.793 ICC with a statistically significant p-value (<0.05) indicates a high degree of agreement among the evaluators. The 95% Confidence Interval of $[0.69, 0.86]$ further supports the reliability across the evaluated items.

To compare the performance of the metrics with humans on these 75 source-translation pairs, we report the average rank correlation coefficients of Kendall's τ (KENDALL, 1938) and Spearman's ρ (Spearman, 1904) in Table 5. In addition, we also meta-evaluated in terms of pair-wise accuracy (Mathur et al., 2020), which measures the proportion of times the metric correctly reflects the relative ordering of items as compared to human judgments across all item pairs. We compared the human ranks and metric ranks of HC, MF, and MC in pairs as mentioned in Table 6.

Type	ICC3k
Description	Average fixed raters
ICC	0.792572
p-value	3.537376e-16
CI95%	[0.69, 0.86]

Table 4: Intra-class Correlation Coefficient (ICC)

Metrics	**Kendall's τ**	**Spearman's ρ**
MLCoT_GPT	**0.3** (1)	**0.329** (1)
TEAM_GPT	**0.227** (2)	**0.233** (2)
TEAM_Aya23	**0.145** (3)	**0.148** (3)
TEAM_Gemma	0.081 (4)	0.096 (4)
CoT_GPT*	0.002 (5)	0.014 (5)
TEAM_Llama	-0.064 (6)	-0.040 (6)
CometKiwi*	-0.118 (7)	-0.089 (7)

Table 5: Agreement with humans in terms of Kendall's τ and Spearmans's ρ average rank correlation coefficient. Ranks are mentioned in brackets. Top three ranking accuracies are highlighted in bold. Baselines are marked with asterisk(*).

Metrics	**HC v/s MF**	**MC v/s MF**	**HC v/s MC**
TEAM_GPT	**0.64** (1)	**0.36** (1)	**0.36** (3)
MLCoT_GPT	**0.6** (2)	**0.24** (3)	**0.52** (1)
TEAM_Gemma	**0.6** (2)	0.2 (4)	**0.36** (3)
CometKiwi*	0.4 (3)	**0.32** (2)	0.32 (4)
TEAM_Aya23	0.4 (3)	**0.32** (2)	0.28 (5)
CoT_GPT*	0.32 (4)	**0.24** (3)	**0.44** (2)
TEAM_Llama3.1	0.32 (4)	**0.36** (1)	0.32 (4)

Table 6: Pairwise Accuracy of the metrics with human assessments for en-hi subtitles; with rank mentioned in brackets. Top three ranking accuracies are highlighted in bold. Baselines are marked with asterisk(*).

5.3 Discussion

Multi-Agent and Multi-Level CoT Approaches Outperform Baselines: Our proposed methods using GPT4-o-mini, (MLCoT_GPT and TEAM_GPT) consistently demonstrated superior performance when compared to CometKiwi and CoT_GPT, in terms of preference rankings (Table 3), rank correlations (Table 5) and pairwise accuracies (Table 6). Specifically, TEAM_GPT achieved the highest ranking in HC vs. MF (0.64) and MC vs. MF (0.36) pairwise comparisons, indicating a strong alignment with human preferences for favoring creative subtitles over formal translations. MLCoT_GPT also ranked highly, showing consistency in favoring creative translations and emphasizing the potential of our methods for more accurate translation evaluation of movie subtitles.

Disagreement of Baselines with Humans: The Kendall's τ and Spearman's ρ average rank correlation coefficients (Table 5) highlight the limitations of baseline methods. The negative correlations observed for CometKiwi and CoT_GPT indicate weaker alignment with human judgments. In contrast, MLCoT_GPT achieved the highest correlation with human ranks ($\tau = 0.3$, $\rho = 0.329$), followed by TEAM_GPT, demonstrating that our proposed methods offer a more reliable and consistent assessment of subtitle translations compared to traditional metrics.

Advancing Translation Evaluation with Structured Reasoning and Multi-Agents: The integration of Multi-Agents and structured reasoning with large language models (LLMs) represents a promising approach to translation evaluation, particularly in challenging domains and resource-constrained environments where gold references or human judgments, may be limited. **Our findings highlight the potential of LLMs to act as human-like evaluators**, marking a significant advancement in the field of translation evaluation.

Agent Benefits with Retrieval-Augmented Generation (RAG) Pipeline: As we see in Table 3 the ability to retrieve relevant examples in real time enabled the agent to make better informed decisions, resulting in improved alignment with human judgements.

6 Conclusion

Translating subtitles is a complex task, and to streamline the process, we proposed novel evaluation methods that can be seamlessly integrated into the subtitle post-editing pipeline to help post-editors identify the most suitable translations. By leveraging the broader context and reasoning abilities of LLMs, we explored several evaluation methods, including Chain-of-Thought (CoT), Multi-level Chain-of-Thought (MLCoT), and Multi-agent evaluation (with and without the Retrieval-Augmented Generation (RAG) pipeline).

We compared our methods with the state-of-the-art CometKiwi metric and observed that CometKiwi favored formal translations over more creative ones. Among our proposed approaches, the multi-agent evaluation method outperformed others in preference ranking and agreement with human evaluators. Additionally, we tested our approach using publicly available LLMs, such as Gemma, Aya, and Llama 3.1. Our results show that the multi-agent evaluation, even with smaller LLMs, outperformed the baseline, highlighting the effectiveness and adaptability of our method.

As a part of future work, we plan to explore LLM's reasoning capabilities further with finer prompts and include other language families, enhancing the robustness and applicability of TEAM across diverse linguistic contexts.

Limitations

Trained Evaluation Metric: A limitation of our work is that we focus on inference rather than training the models, primarily due to the unavailability of movie-subtitles translation evaluation dataset. This hindered our ability to fine-tune or train the models for more accurate and context-sensitive translations. A possible future direction could be to develop an online evaluation metric which adapts to the human evaluator's post-edits on-the-fly, automatically adjusting to the fine-grained nuances specific to the movie instance being translated.

Possible LLM Bias: As we use LLMs for the evaluation system, possible biases in training data and model assumptions may influence outcomes, which might lead to skewed assessments.

Inference Time: Our approaches, MLCoT and TEAM, take longer time to evaluate compared to the baselines, COMETKIWI and CoT_GPT as mentioned in Table 11 in Appendix A.3. However, this increased inference time comes with a trade-off, potentially offering higher agreement with human assessments.

Language Limitations: Although the proposed evaluation methods perform well as shown, they may still face challenges when applied to low-resource languages, which are underrepresented in large-scale LLM training data. This can result in poor performance or translation inaccuracies when handling these languages.

Ethics Statement

Dataset: We sourced our dataset from Opensubtitles, ensuring that all sensitive or personally identifiable information has been removed. However, it is important to note that the movie subtitles may exhibit biases associated with particular genres, cultures, or regions. These biases should be considered when interpreting the results of any analyses conducted using this dataset.

LLMs: Our experiments involve the use of large language models (LLMs), which may carry biases based on the data they were trained on. These biases can potentially affect the justifications generated by our methods. To mitigate this, all model-generated justifications should undergo manual review to ensure accuracy, fairness, and alignment with ethical standards.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Ana Guerberof Arenas and Antonio Toral. 2022. CREAMT: Creativity and narrative engagement of literary texts translated by translators and NMT. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 357–358, Ghent, Belgium. European Association for Machine Translation.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.

Sheila C. M. de Sousa, Wilker Aziz, and Lucia Specia. 2011. Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 97–103, Hissar, Bulgaria. Association for Computational Linguistics.

Sara Court and Micha Elsner. 2024. Shortcomings of LLMs for low-resource translation: Retrieval and understanding are both the problem. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1332–1354, Miami, Florida, USA. Association for Computational Linguistics.

Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. Improving LLM abilities in idiomatic translation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 175–181, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Team Gemma. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

H. Gottlieb. 1997. *Subtitles, Translation & Idioms*. Center for translation studies University of Copenhagen.

M. G. KENDALL. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Langauge Technology (HLT-NAACL 2003)*, pages 48–54. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Emmy Liu, Aditi Chaudhary, and Graham Neubig. 2023. Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15095–15111, Singapore. Association for Computational Linguistics.

Team LLama. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816, Bangkok, Thailand. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Evgeny Matusov, Patrick Wilken, and Yota Georgakopoulou. 2019. Customizing neural machine translation for subtitling. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy. Association for Computational Linguistics.

Ananya Mukherjee, Hema Ala, Manish Shrivastava, and Dipti Misra Sharma. 2020. Mee: An automatic metric for evaluation using embeddings for machine translation. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 292–299. IEEE.

Ananya Mukherjee and Manish Shrivastava. 2023. Mee4 and xlsim : Iiit hyd's submissions' for wmt23 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 798–803, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*.

Jan Pedersen. 2005. How is culture rendered in subtitles. In *MuTra 2005–Challenges of multidimensional translation: Conference proceedings*, volume 18. Citeseer.

Zoë Pettit. 2009. *3: Connecting Cultures: Cultural Transfer in Subtitling and Dubbing*, pages 44–57. Multilingual Matters, Bristol, Blue Ridge Summit.

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

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

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and

André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ayako Sato, Kyotaro Nakajima, Hwichan Kim, Zhousi Chen, and Mamoru Komachi. 2024. TMU-HIT‘s submission for the WMT24 quality estimation shared task: Is GPT-4 a good evaluator for machine translation? In *Proceedings of the Ninth Conference on Machine Translation*, pages 529–534, Miami, Florida, USA. Association for Computational Linguistics.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

C. Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems.*

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS ’22, Red Hook, NY, USA. Curran Associates Inc.

Minghao Wu, Jiahao Xu, and Longyue Wang. 2024. TransAgents: Build your translation company with language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 131–141, Miami, Florida, USA. Association for Computational Linguistics.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni: Language models can hear, talk while thinking in streaming. *Preprint*, arXiv:2408.16725.

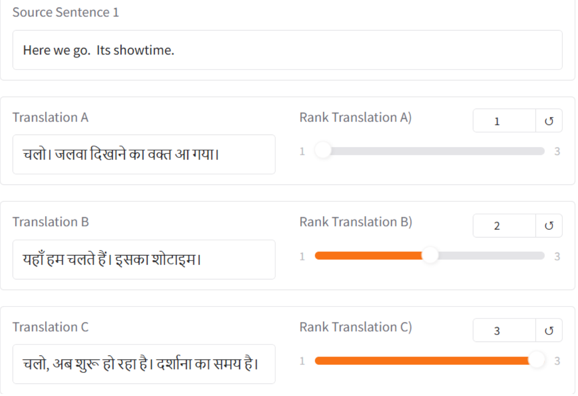
Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations.*

A Appendix

A.1 Human Evaluation

For the human evaluation, three graduate students volunteered to assess the En-Hindi movie subtitle translations based on their preferences. All three evaluators were native Hindi speakers with proficiency in English. The evaluators were blinded to the source-system of the translations. Using the interface as shown in Figure 4, they were asked to rank the translations on a scale of 1 to 3, with 1 representing the best and 3 representing the worst.



Source Sentence 1

Here we go. Its showtime.

Translation A

चलो। जलवा दिखाने का वक्त आ गया।

Rank Translation A) 1

Translation B

यहाँ हम चलते हैं। इसका शोटाइम।

Rank Translation B) 2

Translation C

चलो, अब शुरू हो रहा है। दर्शाना का समय है।

Rank Translation C) 3

Figure 4: Screenshot depicting the human evaluation interface.

A.2 COT and MLCoT Prompts

Table 7 presents the Chain-of-Thought prompt, which is a refined version of the GEMBA prompt, used in our experiments as a baseline for adapting it to the entertainment domain. Tables 8 and 9 display the prompts for Senior and Junior Reviewers, respectively, in the MLCoT experiments.

A.3 LLMs

The details of the Large Language Models, including their names, sizes, and other hyperparameters used in our work, are reported in Table 10.

A.4 Metrics

Table 12 provides the signatures and source code details of the automatic evaluation metrics used in our study.

A.5 Machine Generated Creative Translations

Using the prompt outlined in Table 13, we prompted the Large Language Model (Gemma) to generate Machine-Generated Creative Translations of subtitles for our test set.

A.6 Prompts used in TEAM approach

The prompts used in our Multi-Agentic approach, TEAM, are mentioned in Table 14, 15, 16, 17 and 18.

CoT_prompt = f"""You are a Linguistic Quality Assurance expert in specializing in entertainment content (movies, TV shows) in {src_lng} and {tgt_lng}. Your task is to evaluate the machine generated translations in terms of accuracy, naturalness, tone, intent, style, and cultural appropriateness focusing on entertainment-specific aspects such as idiomatic expressions, humor, cultural references, and context.
Identify translation errors in categories such as {error_ctg}.
Specify the error span in the target sentence and classify severity (major,minor).
Assign a final evaluation score (0-100).
Do not penalize transliteration where appropriate, but penalize unnatural or flawed translations that disrupt fluency and context.
Strictly provide the output in a json format containing the source, translation, error category, severity and final sentence score.
Source Text: {src_txt} Target Text: {tgt_txt} """

Table 7: Chain-of-thought Prompt

sr_prompt = f"""You are a senior Linguistic Quality Assurance expert in specializing in entertainment content (movies, TV shows) in {src_lng} and {tgt_lng}.Your task is to reverify the junior reviewer's evaluation of machine generated translations in terms of {criteria} focusing on entertainment-specific aspects such as idiomatic expressions, humor, cultural references, and context.
Jr.Reviewer evaluations: {jr_evaluation}.
Jr.Reviewer has specified the error spans and identified errors in categories such as {error_ctg}.
Re-assess the jr.reviewer's evaluations and evaluate the {tgt_lng} translations. If necessary, modify the assessment accordingly.
Assign a final evaluation score (0-100).
Strictly provide only the final sentence score output in a json format (score:).
Source Sentence:{src} Target Sentence:{tgt}"""

Table 8: Prompt to Sr.Reviewer in MLCoT Prompting Technique

jr_prompt = f"""You are a junior Linguistic Quality Assurance expert in specializing in entertainment content (movies, TV shows) in {src_lng} and {tgt_lng}. Your task is to evaluate the machine generated translations in terms of {criteria} focusing on entertainment-specific aspects such as idiomatic expressions, humor, cultural references, and context. Specify the error span and identify translation errors in categories such as {error_ctg}.
Strictly provide the output in a json format containing the evaluations i.e., error spans and error category.
Source Sentence:{src} Target Sentence:{tgt}"""

Table 9: Prompt to Jr.Reviewer in MLCoT Prompting Technique

Model	**Variant**	**Size**	**Temperature**	**Maxtokens**
GPT	gpt-4o-mini	-	0.01	512
Gemma	google/gemma-2-9b-it	9B	0.1	512
Aya23	CohereForAI/aya-23-8B	8B	0.1	512
Llama3.1	meta-llama/Llama-3.1-8B-Instruct	8B	0.1	512

Table 10: Model Details and Hyper-parameters

	Time to run	
Metric	100 samples ↓	1 sample ↓
CometKiwi	**108.6**	**1.08**
CoT_GPT	190.01	1.9
MLCoT_GPT	787.89	7.87
TEAM_GPT (w/o RAG)	2622.37	26.22
TEAM_GPT	2837.75	28.37
TEAM_Gemma	5206.91	52.06
TEAM_Aya23	4576.77	45.76
TEAM_Llama3.1	4381.09	43.81

Table 11: Metric-wise Inference time in seconds.

Metric	Signature / Code
BLEU	nrefs:1, case:mixed, eff:no, tok:13a, smooth:exp, version:2
ChrF++	nrefs:1, case:mixed, eff:yes, nc:6, nw:2, space:no, version:2
BERTScore	`https://pypi.org/project/bert-score/`
MEE4	`https://github.com/AnanyaCoder/WMT22Submission`
Comet	`https://huggingface.co/spaces/evaluate-metric/comet`
CometKiwi	`https://huggingface.co/Unbabel/wmt22-cometkiwi-da`

Table 12: Signatures and Source Code Details of Automatic Evaluation Metrics

```
prompt = f"""You are an Expert Movie Dialogue Translator.
Translate Below English Dialogue to Hindi in an informal way, following below Instructions:
Translate the English dialogues by understanding the meaning, making it engaging and interesting for Hindi Speakers
Translation Should be Accurate Only Provide Final Translation as output and do not provide any explanations.
English Dialogue: {sent}"""
```

Table 13: Prompt to generate MC (Machine Generated Creative Translations)

```
prompt = f"""You are a {role} Linguistic Quality Assurance expert in specializing in entertainment content (movies, TV shows)
in {src_lng} and {tgt_lng}. Your task is to evaluate the machine generated translations focusing on entertainment-specific aspects
such as idiomatic expressions, humor, cultural references, and context.
Evaluate and rate the translation [0-100].
Strictly provide the output in a json format (no code block or additional characters) containing the score (score:).
Source Sentence:{src} Target Sentence:{tgt}"""
if(role == 'senior'):
prompt = prompt + f"""Some Examples: {examples}"""
```

Table 14: Initial Prompt to Jr.Reviewer and Sr.Reviewer in TEAM. Senior reviewer has access to relevant examples retrieved from the Knowledge Base.

```
jr_prompt = f"""You are a junior Linguistic Quality Assurance expert in specializing in entertainment content (movies, TV shows)
in {src_lng} and {tgt_lng}. Your task is to evaluate the machine generated translations focusing on entertainment-specific aspects
such as idiomatic expressions, humor, cultural references, and context.
Consider your previous evaluation score:{jr_response}, Critique Errors:{critique_Errors} and COMET-Kiwi score:{comet_score}.
Re-evaluate the translation [0-100]. Strictly provide the output in a json format (no code block or additional characters)
containing the 'score:' and 'justification:'.
Source Sentence:{src} Target Sentence:{tgt}"""
```

Table 15: Re-Prompt to Junior Reviewer to Adjust based on previous judgement in TEAM

```
prompt = f"""You are a {role} Linguistic Quality Assurance expert in specializing in entertainment content (movies, TV shows)
in {src_lng} and {tgt_lng}. Your task is to evaluate the machine generated translations focusing on entertainment-specific aspects
such as idiomatic expressions, humor, cultural references, and context.
Consider your previous evaluation history, having <source,translation,score & justification>: {history},
Re-evaluate the translation [0-100]. Strictly provide the output in a json format (no code block or additional characters)
containing the 'score:' and 'justification:'."""
```

Table 16: Prompt to 're-assess' in TEAM.

```
sr_prompt = f"""You are a senior Linguistic Quality Assurance expert in specializing in entertainment content (movies, TV shows)
in {src_lng} and {tgt_lng}. Your task is to evaluate the machine generated translations focusing on entertainment-specific aspects
such as idiomatic expressions, humor, cultural references, and context. Previously you have assigned a score of {sr_history[0]}.
Based on your junior reviewers analysis, modify your previous score accordingly (iff necessary)
Junior Reviewer’s analysis: score:{sr_history[1]} and justifications:{sr_history[2]}.
Strictly provide the output in a json format (no code block or additional characters) containing the ‘score:’ and ‘justification:’.
Source Sentence:{source} Target Sentence:{translation}
Some Examples: {rag_examples_with_refined_retrieval}"""
```

Table 17: Re-Prompting Sr.Reviewer with examples extracted by refined retrieval.

```
Q_prompt = f"""You are a Linguistic Quality Assurance expert in specializing in entertainment content (movies, TV shows)
in {src_lng} and {tgt_lng}. Your task is to evaluate the machine generated translations focusing on entertainment-specific
aspects such as idiomatic expressions, humor, cultural references, and context.
Identify the translation errors in terms of {error_ctg}.
Strictly provide the output in a json format (no code block or additional characters) containing the ‘Errors:’ and ‘Severity’.
Source Sentence:{src} Target Sentence:{tgt}"""
```

Table 18: Prompt to Critic to identify errors in TEAM.