# Title

Multi-Perspective Coherence Evaluation (MPCE): A Novel Framework for Assessing Document-Level Machine Translation Quality

# Problem Statement

Current machine translation (MT) evaluation methods often fail to capture the overall coherence and logical flow of translations, especially for longer, more complex texts. Most metrics focus on sentence-level or short-span evaluations, with limited consideration of document-level coherence. This limitation hinders our ability to accurately assess and improve the quality of MT systems for real-world applications that require coherent, contextually appropriate translations of longer documents.

# Motivation

Existing MT evaluation methods, such as BLEU, METEOR, and even more recent neural-based metrics, primarily focus on local, sentence-level assessments. While these metrics are valuable, they often miss crucial aspects of document-level coherence, such as consistent terminology use, logical flow, and overall readability. Our proposed Multi-Perspective Coherence Evaluation (MPCE) framework draws inspiration from multi-agent systems and cognitive science theories of perspective-taking. By simulating diverse reader experiences, MPCE aims to provide a more comprehensive and nuanced evaluation of translation quality that better aligns with human judgments of coherence and readability.

# Proposed Method

The MPCE framework consists of the following key components: 2. Multi-faceted Evaluation Process: For each translation, MPCE deploys these agents to: a) Generate comprehension questions at various levels of granularity b) Attempt to answer these questions based solely on the translation c) Identify logical inconsistencies or coherence breaks d) Provide an overall coherence rating

- Reader Agent Ensemble: We create a diverse set of 'reader agents', each implemented as a fine-tuned language model with carefully crafted prompts to simulate different reader types (e.g., domain expert, casual reader, critical analyst).
- Multi-faceted Evaluation Process: For each translation, MPCE deploys these agents to:
- Graph-based Aggregation: We use a graph neural network to aggregate the multi-perspective insights, where nodes represent different aspects of coherence and edges capture inter-dependencies between perspectives.
- Holistic Coherence Score: The framework produces a nuanced, holistic coherence score that considers diverse reading experiences.

# Step-by-Step Experiment Plan

## Step 1: Data Preparation

Select document-level MT datasets for evaluation. We will use the WMT20 news translation task datasets for English-German and English-Chinese language pairs, as they provide document-level translations and human judgments.

## Step 2: Baseline Metrics Implementation

Implement and run baseline metrics including BLEU, METEOR, BERTScore, and COMET on the selected datasets.

## Step 3: Reader Agent Development

Fine-tune five distinct reader agent models using GPT-

- Domain Expert: Prompt: 'You are an expert in the field. Analyze the text critically for accuracy and depth.'
- Casual Reader: Prompt: 'You are a general reader. Focus on overall clarity and engagement of the text.'
- Language Learner: Prompt: 'You are learning this language. Pay attention to vocabulary usage and sentence structures.'
- Critical Analyst: Prompt: 'You are a critical thinker. Identify logical inconsistencies and evaluate the argument structure.'
- Coherence Specialist: Prompt: 'You are an expert in textual coherence. Assess the logical flow and connections between ideas.'

## Step 4: Question Generation

For each translated document, prompt each reader agent to generate 3-5 comprehension questions at different levels (e.g., factual, inferential, evaluative). Example prompt: 'Based on the given translated text, generate 3-5 diverse comprehension questions that assess understanding at different levels.'

## Step 5: Question Answering

Prompt each reader agent to answer the generated questions based solely on the translated text. Example prompt: 'Answer the following questions based only on the information provided in the translated text.'

## Step 6: Coherence Assessment

Prompt each reader agent to identify logical inconsistencies, coherence breaks, and provide an overall coherence rating (1-10 scale). Example prompt: 'Analyze the text for logical inconsistencies and coherence breaks. Provide specific examples and an overall coherence rating from 1 to 10.'

## Step 7: Graph Neural Network Implementation

Implement a Graph Neural Network (GNN) using PyTorch Geometric library. Nodes represent coherence aspects (e.g., logical flow, terminology consistency) and reader perspectives. Edge weights are determined by the agreement between reader agents on specific aspects.

## Step 8: MPCE Score Computation

Aggregate the multi-perspective insights using the GNN to produce the final MPCE score. The node features will be initialized with the coherence ratings and the number of identified inconsistencies from each reader agent.

## Step 9: Evaluation and Comparison

Compare MPCE scores with baseline metrics and human judgments using Pearson and Spearman correlations. Conduct statistical significance tests to validate improvements.

## Step 10: Ablation Studies

Perform ablation studies by removing different reader agent types and varying the number of agents to assess their individual contributions to the final MPCE score.

# Test Case Examples

## Baseline Prompt Input

Translate the following German text to English: 'Die Klimakrise ist eine der größten Herausforderungen unserer Zeit. Wissenschaftler warnen seit Jahren vor den Folgen der globalen Erwärmung. Trotzdem scheinen viele Regierungen die Dringlichkeit des Problems zu unterschätzen. Es ist wichtig, dass wir jetzt handeln, um die schlimmsten Auswirkungen zu verhindern.'

## Baseline Prompt Expected Output

The climate crisis is one of the biggest challenges of our time. Scientists have been warning about the consequences of global warming for years. Nevertheless, many governments seem to underestimate the urgency of the problem. It is important that we act now to prevent the worst impacts.

## Proposed Prompt Input (Domain Expert)

You are an expert in climate science. Analyze the following translated text critically for accuracy and depth: 'The climate crisis is one of the biggest challenges of our time. Scientists have been warning about the consequences of global warming for years. Nevertheless, many governments seem to underestimate the urgency of the problem. It is important that we act now to prevent the worst impacts.' Generate 3-5 comprehension questions, answer them based solely on the text, identify any logical inconsistencies or coherence breaks, and provide an overall coherence rating from 1 to 10.

## Proposed Prompt Expected Output (Domain Expert)

Comprehension Questions: 4. What does the text suggest we should do? Answers: 4. The text suggests that we should act now to prevent the worst impacts. Logical Inconsistencies/Coherence Breaks: No major logical inconsistencies found. However, the text lacks specific examples or data to support its claims, which could improve its depth and coherence. Overall Coherence Rating: 7/10 The text presents a clear message about the climate crisis, but could benefit from more specific information and a stronger logical flow between sentences.

- What is described as one of the biggest challenges of our time?
- How long have scientists been warning about global warming consequences?
- What is the attitude of many governments towards the problem?
- What does the text suggest we should do?
- The climate crisis is described as one of the biggest challenges of our time.
- Scientists have been warning about the consequences of global warming for years.
- Many governments seem to underestimate the urgency of the problem.
- The text suggests that we should act now to prevent the worst impacts.

## Explanation

The MPCE framework allows for a more nuanced evaluation of the translation by simulating different reader perspectives. In this case, the domain expert provides insights into the accuracy and depth of the content, which goes beyond surface-level linguistic comparisons offered by traditional metrics. This multi-faceted approach captures aspects of coherence and logical flow that are crucial for document-level translation quality but often missed by existing methods.

# Fallback Plan

If the proposed MPCE framework does not show significant improvements over baseline metrics, we will conduct a thorough analysis to understand the limitations. This may include: 1) Examining the quality and diversity of questions generated by reader agents to ensure they capture different aspects of coherence. 2) Analyzing the agreement between different reader agents to identify potential biases or inconsistencies in their evaluations. 3) Investigating the effectiveness of the GNN in capturing inter-dependencies between different coherence aspects. Based on these analyses, we may refine the reader agent prompts, experiment with alternative aggregation methods (e.g., attention mechanisms instead of GNN), or incorporate additional coherence-specific features into the evaluation process. If these refinements do not yield substantial improvements, we will pivot the project towards an in-depth analysis of why capturing document-level coherence in MT evaluation is challenging, potentially uncovering valuable insights for future research directions in this area.

Ranking Score: 6