MT QE Thesis Report By Satya Iyer.docx

by Satya Satya Padmanabhan Iyer

Submission date: 25-May-2025 06:08PM (UTC+0100)
Submission ID: 260385906
File name: MT_QE_Thesis_Report_By_Satya_Iyer.docx (2.99M)
Word count: 19314
Character count: 127183

EVALUATING MACHINE TRANSLATION WITH LARGE LANGUAGE MODELS: A CASE STUDY FOR ENGLISH-HINDI AND -MARATHI

SATYA PADMANABHAN IYER

Thesis Report

MAY 2025

DEDICATION

This final thesis report is dedicated to all those who have inspired and supported me on this journey. To my family, whose unwavering love and encouragement have been my anchor through the highs and lows of this endeavour. To my friends and mentors, whose guidance and wisdom has enriched my understanding and fuelled my passion for research. And to all the individuals who have shared their knowledge and expertise, shaping the path of this project. Your belief in me has been a driving force, and I dedicate this work to each of you with heartfelt appreciation.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all those who have contributed to the development and progress of this Final Thesis research report. Special thanks to my supervisor Dr. Diptesh Kanojia for their guidance, support and invaluable feedback throughout this process. I am also grateful to Liverpool John Moore University, Liverpool, UK and Upgrad Education, India for providing resources and facilities essential for conducting this research. Additionally, I extend my appreciation to my colleagues and peers for their insightful discussions and encouragements. Their input has been instrumental in shaping the direction of this research. Lastly, I would like to acknowledge the unwavering support of my family and friends, whose encouragement has been a constant source of motivation.

ABSTRACT

Machine Translation (MT) evaluation using Quality Estimation (QE) methods traditionally relies on encoder-based models to perform a regression task that predicts translation quality on a scale of 0–100 [or 0-5]. This process is essential for assessing MT output without requiring human-labelled reference translations. In our work, we leverage existing datasets from the Workshop on Machine Translation (WMT), which contain key information such as source sentences, machine-generated translations, and human reference translations. These elements are crucial for evaluating MT quality, whereas individual translation errors may not always be significant.

Conventional MT evaluation metrics, such as BLEU, TER, and chRF, primarily rely on lexical comparisons between the MT output and a single reference translation. However, these approaches often struggle to capture semantic nuances, idiomatic expressions, and domain-specific variations, leading to limitations in assessing translation quality comprehensively. To address these shortcomings, our research focuses on developing computational models that utilize Large Language Models (LLMs) within the Generative AI (GenAI) paradigm for enhanced QE performance.

Our approach involves fine-tuning state-of-the-art LLMs, such as LLaMA-3.2-3B-Instruct Mistral-7B-Instruct-v0.2 and gemma-1.1-4b-it using a diverse multilingual corpus enriched with high-quality translation-related instructions. By incorporating instruction tuning, we aim to create models that can generalize effectively across multiple translation tasks and language pairs. We will systematically compare our results against established baseline approaches, including TransQuest, COMET, and traditional statistical metrics such as Spearman, Pearson and Mean Absolute Error (MAE), to validate the effectiveness of our method. We also used Prompting Techniques like Zero Shot and Few Shot Prompting to analyze the result.

The ultimate objective of this research is to develop an open-source Quality Estimation (QE) model that not only surpasses existing state-of-the-art approaches but also demonstrates competitive performance with high-capacity decoder-only Large Language Models (LLMs) in the task of translation quality estimation. The performance of the proposed model will be rigorously evaluated using standard QE benchmarks and compared against established baselines, including both traditional QE systems and instruction-tuned decoder-based LLMs.

TABLE OF FIGURES

TABLE OF FLOWCHARTS

LIST OF TABLES

LIST OF ABBREVIATIONS

MT	Machine Translation
QE	Quality Estimation
LLM(s)	Large Language Model(s)
QuEst	Quality Estimation Toolkit (early QE system)
QuEst++	An improved version of QuEst
MARMOT	Not explicitly expanded (QE tool using CRFs; often treated as a name)
POSTECH	Pohang University of Science and Technology (often refers to QE models developed there)
BERT	Bidirectional Encoder Representations from Transformers
XLM	Cross-lingual Language Model
XLM-R	Cross-lingual Language Model – RoBERTa
mDistilBERT	Multilingual Distilled BERT
OpenKiwi	Open-source Toolkit for Quality Estimation
deepQuest	Deep Learning Framework for QE
TransQuest	Transformer-based Quality Estimation toolkit
COMET	Crosslingual Optimized Metric for Evaluation of Translation
DA	Direct Assessment
LLaMA	Large Language Model Meta AI
CoT	Chain of Thought (reasoning)
en–hi	English–Hindi
en–mr	English–Marathi

CRF(s)	Conditional Random Field(s)
QUETCH	QE model introduced in early predictor-estimator frameworks
NuQE	Neural Quality Estimation
mBERT	Multilingual BERT
TLM	Translation Language Modeling (variant of XLM)
Name	Description
Word2Vec	Word Embedding Model
GloVe	Global Vectors for Word Representation
QuEst / QuEst++	Traditional QE models
QUETCH	Neural QE model
NuQE	Neural QE Predictor-Estimator
MARMOT	QE model using Conditional Random Fields
POSTECH	QE model using Predictor-Estimator architecture
BERT	Bidirectional Encoder Representations from Transformers
mBERT	Multilingual BERT
TLM	Translation Language Modeling
XLM	Cross-lingual Language Model
XLM-R	XLM-RoBERTa (Cross-lingual RoBERTa)
mDistilBERT	Multilingual Distilled BERT
deepQuest	Open-source QE framework
OpenKiwi	Open-source QE framework
TransQuest	Transformer-based QE model using XLM-R
COMET	Cross-lingual Optimized Metric for Evaluation of Translation
GEMBA	GPT-4 based QE evaluation approach
GPT-4	Generative Pre-trained Transformer version 4

LLaMA-2	Large Language Model by Meta AI (2nd generation)
LLaMA-3.2-3B-Instruct	Instruction-tuned version of LLaMA 3.2 with 3B parameters
Mistral-7B-Instruct-v0.2	Instruction-tuned version of Mistral with 7B parameters
Gemma-1.1-4B-IT	Instruction-tuned Gemma model with 4B parameters
BERTScore	Evaluation metric using BERT embeddings
WMT	Workshop on Machine Translation (shared task platform)

Table of Contents

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

The rapid advancements in Natural Language Processing (NLP) have been largely driven by the emergence of Large Language Models (LLMs), which have demonstrated state-of-the-art performance across various tasks, including question answering, text summarization, information retrieval, and, most notably, machine translation (MT) (Kocmi and Federmann, 2023) With the increasing availability of large-scale multilingual datasets and computational power, LLMs have significantly enhanced the capabilities of MT systems, enabling more accurate and context-aware translations.

Traditionally, MT quality assessment has relied on evaluation metrics such as BLEU (Papineni *et al.*, 2002), BLEURT (Sellam, Das and Parikh, 2020), and BERTScore (Zhang *et al.*, 2024) which compare machine-generated translations against human reference translations. While these metrics have been widely used, they often struggle to capture nuanced linguistic phenomena such as idiomatic expressions, paraphrasing, and domain-specific terminology. Moreover, in real-world scenarios, reference translations are not always available, making it challenging to apply these traditional evaluation techniques effectively.

To address this limitation, Quality Estimation (QE) methods have been developed to assess MT output without requiring reference translations. QE techniques typically fine-tune multilingual pre-trained models on human evaluation datasets, such as Direct Assessment (DA) scores, which provide a standardized measure of translation quality on a scale of 0 to 100 (Graham *et al.*, 2020; Kanojia *et al.*, 2021; Zerva *et al.*, 2022a) These scores are often transformed into z-scores to normalize variations and enhance the training of QE models. Recent studies have explored the potential of prompting LLMs to directly assign translation quality scores, showing promising results in automatic MT evaluation (Kocmi and Federmann, 2023)

Two notable advancements in Machine Translation Quality Estimation are TransQuest (Ranasinghe, Orăsan and Mitkov, 2020)and COMET,(Chimoto and Bassett, 2022) each offering unique strengths. TransQuest is a lightweight, multilingual framework designed for

both word-level and sentence-level QE. It significantly reduces computational requirements while maintaining high performance, especially in low-resource settings, and has outperformed earlier QE models across 15 language pairs, winning multiple WMT shared tasks. In contrast, COMET leverages pretrained multilingual encoders and human assessment data to predict sentence-level quality with high accuracy. By fine-tuning on Direct Assessment scores, COMET effectively captures both semantic adequacy and fluency, achieving top rankings in recent WMT evaluations(Specia *et al.*, 2020; Heafield, Zhu and Grundkiewicz, 2021a; Zerva *et al.*, 2022b; Bhattacharyya *et al.*, 2023a). Together, these models represent complementary approaches: TransQuest emphasizes efficiency and adaptability, while COMET excels in leveraging human-annotated quality signals for enhanced prediction accuracy.

Multilingual NLP encompasses a broad spectrum of tasks beyond quality estimation, including automated revision, linguistic error correction, and cross-lingual text generation(Conneau *et al.*, 2020). These tasks require handling and generating text across multiple languages while preserving meaning, coherence, and stylistic consistency. The increasing demand for high-quality translations across diverse domains has led to a paradigm shift from traditional task-specific models toward versatile, large-scale LLMs, which have achieved cutting-edge results in multiple recent WMT shared challenges.

However, despite their potential, publicly available LLMs still face limitations in multilingual translation tasks, particularly when compared to proprietary models like GPT-4. Many open-source LLMs are predominantly optimized for English-language tasks, which restricts their effectiveness in multilingual settings. Consequently, methods that leverage these models often struggle to achieve competitive performance unless they are fine-tuned on specific objectives. This disparity highlights the need for further research in adapting open-source LLMs for high-quality multilingual translation assessment, ensuring they can match or surpass the capabilities of proprietary systems.

To evaluate the performance of Quality Estimation models, it is essential to compare the predicted quality scores with human-annotated reference scores, such as Direct Assessment (DA) ratings. Here, correlation metrics serve as key indicators of how well a model captures translation quality as perceived by human evaluators. Spearman's rank correlation coefficient assesses the monotonic relationship between the predicted and actual scores, focusing on

whether the relative ordering of translations is preserved rather than the exact numerical values. Pearson's correlation coefficient, on the other hand, measures the linear relationship between predicted and reference scores, indicating how closely the prediction trend follows the ground truth. While Spearman is robust to non-linear relationships, Pearson is more sensitive to exact score alignment. Additionally, Mean Absolute Error (MAE) quantifies the average absolute deviation between the predicted scores and human judgments, providing a direct interpretation of prediction accuracy in the original score scale. Together, these metrics offer a comprehensive evaluation framework: Spearman captures consistency in ranking, Pearson reflects the strength of linear trends, and MAE reveals the magnitude of prediction error. Using all three enables us to diagnose different aspects of QE model behaviour, which is especially crucial when models are deployed in high-stakes multilingual environments where both score accuracy and relative ranking are critical.

In this paper, we address this problem by developing multilingual sentence-level QE models which perform competitively in different domains, MT types and language pairs, we propose sentence-level QE as a zero-shot and Few Shot cross lingual transfer task, enabling new avenues of research in which multilingual models can be trained once and then serve a multitude of languages and domains.

The main contributions of this research are the following:

i. We introduce a simple architecture to perform sentence-level quality estimation that predicts the quality of the sentence in the source sentence, target sentence and the gaps in the target sentence.
ii. We explore multilingual, sentence-level quality estimation with the proposed architecture. We show that multilingual models are competitive with bilingual models.
iii. We inspect few-shot and zero-shot sentence-level quality estimation with the bilingual and multilingual models. We report how the source target direction, domain and MT type affect the predictions for a new language pair.
iv. We release the code and the pre-trained models as part of an open-source framework.

1.2 PROBLEM STATEMENT

Machine Translation (MT) quality estimation (QE) plays a critical role in evaluating the accuracy and fluency of translations, especially when human-labelled reference translations are unavailable. Traditional QE models rely on encoder-based architectures and fine-tuned

multilingual pre-trained models, such as XLM-R, which is trained on text from 100 languages. While effective, these models often require extensive training and struggle to generalize across different languages and domains. Additionally, traditional QE approaches face challenges in assessing translation quality when dealing with highly diverse linguistic structures, idiomatic expressions, and domain-specific terminologies.

Our research aims to address these challenges by developing multilingual sentence-level QE models that achieve strong results across various domains, translation styles, and language combinations. Unlike traditional methods, our approach introduces sentence-level QE as a zero-shot and Few Shot cross-lingual adaptation task, enabling multilingual models to be trained once and applied across multiple languages and domains without requiring explicit labelled examples for each language. This significantly enhances scalability and adaptability, opening new avenues for research in multilingual translation quality estimation.

Furthermore, recent studies have demonstrated that increasing model capacity can mitigate the so-called "curse of multilingualism", thereby improving cross-lingual performance. To validate this, our study explores the impact of zero-shot learning (ZSL) and Few-shot learning (FSL) reasoning in the context of QE, leveraging these techniques to enhance translation accuracy and model interpretability. Figure 1 Explains the architectural diagram of Woking style of Machine Translation with Quality Scores.

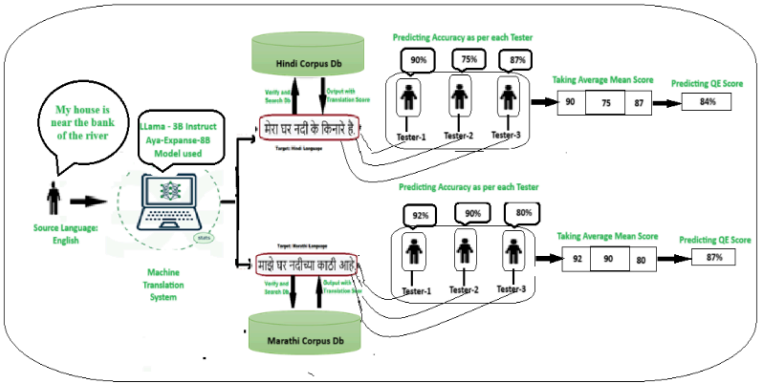


Figure 1: Architecture Design of Machine Translation with Quality Estimation Scores from English-Hindi and English-Marathi

Figure1 depicts a machine translation system that converts an English sentence into Hindi and Marathi, (vice-versa) that verifies the translations against a corpus, and assigns accuracy scores. Now, let's analyse and understand how Zero-shot Learning (ZSL) and Few-shot learning (FSL) reasoning can be applied in this context.

1.2.1 Zero-Shot Translation Methodology:

It refers to a model's ability to translate between languages it has never seen before in training. The model relies on its general language understanding without explicit labelled examples,

Step 1: Input Sentence (English): "My house is near the bank of the river."
(This is an ambiguous sentence because "bank" could mean a financial bank or riverbank.)

Step 2: Machine Translation (Without Prior Hindi/Marathi Training) The model has not been explicitly trained with Hindi/Marathi parallel text but still attempts to generate meaningful translations:

Hindi Translation: "***मेरा घर नदी के किनारे है।***"
(Correctly captures "bank" as "**किनारे**" meaning riverbank.)

Marathi Translation: "**माझे घर नदीच्या काठी आहे।**"
(Also correctly represents riverbank as "**काठी.**")

Step 3: Verification and QE Scoring:

The system checks the accuracy score given by 3 Tester based on sentence level translation. After getting the score it will calculate the Mean z-score and finally the Quality Estimation Score is Predicted

Hindi: 90,75,87 → **Calculating mean gives us QE Score of given sentences is 84%.**
Marathi: 92,90,80 → **Calculating mean gives us QE Score of given sentences is 87%.**

The model correctly translates despite never being trained explicitly on Hindi/Marathi.
It understands context, structure, and word meanings based on its general multilingual knowledge.

1.2.2 Few-shot learning (FSL) Translation Methodology

Few-shot translation refers to a model's ability to generalize to a new language pair after being exposed to a small number of labelled examples during training or inference. Unlike zero-shot learning, the model is provided with limited parallel sentences that help it establish mappings between source and target language structures and semantics.

Step 1: Input Sentence (English):

"The children are playing in the garden."

Step 2: Few-Shot Prompt (Using English–Hindi/Marathi pairs):

The model is given a few translation examples as context before translating the new input.

Examples Provided to the Model:

- ***English:*** **"The dog is sleeping on the floor."**

  Hindi: **"कुत्ता फर्श पर सो रहा है।" / *Marathi:* "कुत्रा जमिनीवर झोपलेला आहे।"**

- ***English:*** **"She is reading a book."**

  Hindi: **"वह किताब पढ़ रही है।" / *Marathi:* "ती पुस्तक वाचत आहे।"**

Now translating the target input...

Hindi Translation: "बच्चे बगीचे में खेल रहे हैं।"

(Correctly reflects plural form "children" and location "garden.")

Marathi Translation: "मुले बागेत खेळत आहेत।"

(Accurately captures tense, subject agreement, and context.)

Step 3: Verification and QE Scoring:

To assess the quality of the few-shot translations, three human evaluators assign sentence-level quality scores. The system then computes the mean score to estimate overall translation quality.

- **Hindi:** 85, 88, 90 → Mean QE Score: **87.7%**
- **Marathi:** 83, 86, 89 → Mean QE Score: **86%**

Conclusion:

Despite only having access to a few supervised translation examples, the model demonstrates robust contextual understanding and accurate grammatical generation. The relatively high QE

scores affirm the effectiveness of few-shot prompting in aiding translation for underrepresented language pairs.

Our contributions are:

Our research introduces a novel approach to sentence-level QE that integrates zero-shot cross-lingual adaptation and reasoning-based translation methodologies to improve MT evaluation. Our key contributions include:

- We investigate the key components required for evaluating translations using LLMs, including source text, reference translations, translation errors, and annotation instructions.
- We analyze the effectiveness of zero-shot prompting and few-shot prompting for machine translation quality assessment.
- Our study systematically compares prompt-based methods with fine-tuned multilingual pre-trained language models (PTLMs) to determine their relative performance in QE. We find that while LLMs offer strong generalization capabilities, they still lag behind fine-tuned PTLMs in certain translation quality tasks.
- Our analysis reveals that reference texts play a crucial role in achieving accurate translation assessments. We emphasize the importance of designing effective prompts for LLM-based QE systems.

By developing multilingual sentence-level QE models, introducing zero-shot and Few shot cross-lingual adaptation, and leveraging reasoning-based translation methodologies, our research aims to bridge the gap between open-source and proprietary LLMs in multilingual QE tasks.

1.3 RESEARCH OBJECTIVE

These research questions guide an in-depth exploration of the evolution, strengths, and potential of modern techniques in machine translation evaluation, with a particular focus on LLMs.

1. How have Quality Estimation (QE) models for Machine Translation (MT) evolved over the past decade, and what are the key advancements in techniques, architectures, and performance improvements?

2. What are the strengths and limitations of early-generation QE models like traditional metrics (e.g., BLEU, TER) compared to more recent advancements such as transformer-based models like BERT and multilingual models like mBERT, XLM, and XLM-RoBERTa?
3. Can LLMs like LLaMA-3.2-3B-Instruct, Mistral-7B-Instruct-v0.2, and Gemma-1.1-4B-IT provide more accurate and efficient translation quality estimation concerning state-of-the-art QE models and traditional evaluation metrics?

1.4 AIMS AND OBJECTIVE

This research aims to evaluate the effectiveness of instruction-tuned large language models (LLMs) such as LLaMA-3.2-3B-Instruct, Mistral-7B-Instruct-v0.2, and Gemma-1.1-4B-IT for Machine Translation Quality Estimation (MT QE). The proposed evaluation is benchmarked against established QE frameworks like TransQuest and COMET, as well as standard statistical metrics—BLEU, TER, and chrF—commonly used for machine translation evaluation. While BLEU focuses on n-gram overlap and suffers from semantic limitations, TER evaluates the edit distance for post-editing but may penalize stylistic variants. chrF, which measures character n-gram similarity, is better suited for morphologically rich languages. This study investigates whether LLM-based QE can provide more context-aware, semantically grounded assessments than these traditional approaches.

The proposed approach uses LLaMA-3.2-3B-Instruct (Dubey *et al.*, 2024), Mistral-7B-Instruct-v0.2 (Jiang *et al.*, 2023)and Gemma-1.1-4b-it(Gemma Team *et al.*, 2024) are LLMs used in natural language processing tasks due to their unique strengths and capabilities. In recent years, instruction-tuned large language models (LLMs) have gained significant attention for their ability to follow user commands, perform multi-task reasoning, and provide coherent, context-aware responses across domains. This section provides a comparative analysis of three prominent open-access instruction-tuned models: LLaMA-3.2-3B-Instruct, Mistral-7B-Instruct-v0.2, and Gemma-1.1-4B-IT.

LLaMA-3.2-3B-Instruct, developed by Meta AI, is a decoder-only transformer model with 3.2 billion parameters, designed for efficient instruction-following. Its lightweight architecture and low-latency inference make it particularly suitable for resource-constrained environments, such as edge devices or single-GPU setups. This is especially advantageous for deploying MT

Quality Estimation (MT QE) solutions for Indic languages like English-Hindi (en-hi) and English-Marathi (en-mr), where infrastructure may be limited.

In contrast, Mistral-7B-Instruct-v0.2, introduced by Mistral AI, is a more powerful 7 billion parameter model featuring grouped-query attention and sliding window positional encodings. These innovations enable high memory efficiency and long-context processing (up to 32K tokens), which is crucial for few-shot and zero-shot prompt-based QE tasks, especially when dealing with long source-target translation pairs in languages like Hindi and Marathi.

Gemma-1.1-4B-IT, developed by Google DeepMind, offers a balanced alternative with 4 billion parameters and strong instruction alignment. It is optimized for producing high-quality, coherent English responses and performs reliably in multilingual scenarios, including translation evaluation for en-hi and en-mr.

All three models have been instruction-tuned for improved generalization and usability across tasks. LLaMA is tuned for concise, prompt-based outputs, ideal for zero-shot QE scenarios involving short or straightforward translations. Mistral is optimized for complex reasoning and long-form evaluation tasks, making it highly suitable for few-shot evaluation of nuanced translations. Gemma occupies a middle ground and shows good coherence and alignment, especially in structured, guided evaluation tasks.

From a deployment and licensing perspective, LLaMA uses a non-commercial license with limited multilingual capabilities. Mistral is open-source under Apache 2.0 and supports moderate multilingual input. Gemma is released under Google's Gemma license, with strong English alignment but growing support for multilingual applications.

In the context of MT QE for en-hi and en-mr, the study leverages these models to evaluate translation quality without reference translations, using prompt-based techniques. Mistral exhibits strong performance in few-shot setups, while LLaMA and Gemma show competitive performance in zero-shot configurations, offering flexibility across diverse task conditions and computational constraints.

Research Objectives

1. To analyze the comparative strengths of LLaMA-3.2-3B-Instruct, Mistral-7B-Instruct-v0.2, and Gemma-1.1-4B-IT across MT QE and general NLP tasks.

2. To evaluate the models' performance in en-hi and en-mr MT QE using both zero-shot and few-shot prompting techniques.
3. To benchmark LLM-based QE outputs against traditional metrics such as BLEU, TER, chrF, and established models like TransQuest and COMET.
4. To assess trade-offs among model size, computational efficiency, multilingual support, context-handling, and translation output quality.
5. To provide practical insights into the applicability of instruction-tuned LLMs for real-world multilingual MT evaluation scenarios, especially in low-resource settings.

1.5 SIGNIFICANCE OF STUDY

This study holds significant potential to advance the evaluation of machine translation (MT) systems, particularly for low- and mid-resource Indian languages such as English-Hindi (en-hi) and English-Marathi (en-mr). While traditional MT quality estimation relies on regression-based encoder models and surface-level metrics like BLEU, TER, and chrF, these methods often fall short in capturing semantic accuracy, contextual coherence, and idiomatic correctness, especially in morphologically rich and syntactically diverse languages like Hindi and Marathi.

By leveraging large language models (LLMs) under the generative AI (GenAI) paradigm, this study introduces a prompt-based MT quality estimation (MT QE) framework using zero-shot and few-shot learning techniques. These allow the models to assess translation quality without explicit training on the evaluation task, offering adaptability to unseen language pairs and domains. This is especially valuable for resource-scarce language pairs like en-mr, where supervised training data is limited.

The study integrates instruction-tuned LLMs—LLaMA-3.2-3B-Instruct, Mistral-7B-Instruct-v0.2, and Gemma-1.1-4B-IT—and assesses their ability to provide accurate, context-aware translation quality judgments. These models are tested across diverse translation scenarios using WMT datasets, which include source sentences, MT outputs, and human references. This enables the comparison of prompt-based model outputs with traditional evaluation metrics and QE models such as TransQuest and COMET.

By addressing both lexical and semantic dimensions of translation quality, the study provides a more nuanced and comprehensive evaluation framework. Furthermore, it highlights the trade-offs between computational cost, scalability, and translation evaluation performance, offering actionable insights for selecting appropriate models in real-world NLP applications involving Indic languages.

Ultimately, this research contributes to bridging the gap between automated evaluation and human-like quality assessment, especially in multilingual and culturally complex environments. It sets a foundation for more inclusive, reliable, and adaptable MT evaluation methods in the context of Indian languages.

1.6 SCOPE OF STUDY

This study focuses on the evaluation of machine translation (MT) systems specifically for English-Hindi (en-hi) and English-Marathi (en-mr) language pairs—representing both a high-resource and a relatively low-resource Indian language. By selecting these pairs, the study aims to capture the distinct challenges of syntactic divergence, morphological richness, and domain-specific nuances often encountered in Indic language translation.

The analysis will utilize benchmark datasets from the WMT shared tasks, which offer high-quality parallel data and human-annotated translation quality scores. This multilingual dataset enables the evaluation of translation systems across two typologically different target languages under comparable source input conditions.

The study will employ a dual evaluation approach:
Traditional metrics including BLEU, TER, and chrF to assess surface-level n-gram overlap and post-editing effort.
Modern quality estimation methods such as TransQuest and COMET, which are capable of capturing deeper semantic fidelity and contextual accuracy without relying solely on reference translations.

The core analysis will benchmark three instruction-tuned large language models—LLaMA-3.2-3B-Instruct, Mistral-7B-Instruct-v0.2, and Gemma-1.1-4B-IT—for their performance in translation quality estimation (MT QE). The models will be evaluated based on their ability to

follow instructions, infer semantic correctness, and generate coherent judgments about translation quality in the en-hi and en-mr contexts.

In addition, the study will examine the computational efficiency, scalability, and alignment quality of these models in practical MT evaluation workflows. Special attention will be paid to how each model handles complex translation phenomena such as idiomatic usage, domain-specific vocabulary, and code-mixed input often present in Indic language scenarios.
In summary, the scope of this research is defined by its emphasis on evaluating MT quality for English to Hindi and Marathi, leveraging a hybrid of lexical and semantic scoring approaches, and assessing the applicability of modern LLMs in handling linguistically diverse and culturally rich language pairs.

1.7 THESIS STRUCTURE

This thesis is structured into six chapters, each addressing a critical aspect of Machine Translation (MT) Quality Estimation (QE).
The **Chapter1 -Introduction** presents the research problem, motivation, and objectives, emphasizing the limitations of traditional translation evaluation and the potential of LLMs and instruction fine-tuning in multilingual QE.
The **Chapter2 -Literature Review** explores existing QE methods, covering statistical, neural, and transformer-based approaches. It reviews traditional metrics like BLEU, TER, BERTScore and examines advancements such as TransQuest, COMET, and XLM-R, highlighting gaps in current research.
The **Chapter3-Research Methodology** details data selection, preprocessing, and model design for sentence-level QE. It introduces Zero-Shot Learning (ZSL), Few-shot Learning (FSL) reasoning, and fine-tuning LLaMA-3.2-3B-Instruct and Mistral-7B-Instruct-v0.2, and Gemma-1.1-4B-IT for multilingual adaptability.
The **Chapter4-Implementation and Analyses** chapter explains the training process, parameter tuning, and model evaluation, detailing how sentence-level QE scoring is applied across languages and how translations are verified against benchmarks.
The **Chapter5-Analyses and Results Discussions** chapter compares the proposed approach with baseline QE models, evaluating improvements in translation accuracy, multilingual adaptability, and fine-tuning techniques through quantitative and qualitative assessments.

The **Chapter6-Conclusion and Future Work** summarizes key findings, addressing limitations in cross-lingual generalization and potential improvements in fine-tuning, low-resource language adaptation, and context-aware evaluation for robust QE.

1.8 SUMMARY

This chapter introduces the growing importance of Quality Estimation (QE) in Machine Translation (MT), driven by advancements in Large Language Models (LLMs). Traditional evaluation metrics like BLEU and BERTScore require reference translations and often fail to capture linguistic nuances. QE models address this by predicting translation quality without needing references. Key approaches include TransQuest, known for its efficiency and multilingual adaptability, and COMET, which uses human-assessed scores to achieve high accuracy.

The research explores sentence-level QE as a zero-shot cross-lingual task using the XLM-R model. This approach allows a single multilingual model to handle multiple language pairs and domains without retraining. A custom architecture incorporating a <GAP> token and a linear prediction layer is proposed to estimate quality scores between 0-100.

Evaluation uses Spearman, Pearson, and MAE metrics to assess correlation and prediction accuracy against human scores. The study also introduces zero-shot, Few Shot and Chain of Thought (CoT) reasoning to improve translation and QE for unseen languages. Demonstrated through English-Hindi and English-Marathi examples, the models show strong contextual understanding even without prior training. Overall, this chapter sets the foundation for building scalable, accurate, and adaptable multilingual QE systems using modern LLM techniques.

CHAPTER-2

LITERATURE REVIEW

2.1 CURRENT KNOWLEDGE

The evolution of Quality Estimation (QE) models for Machine Translation (MT) reflects a continuous evolution in techniques, architectures, and performance improvements over time. These developments can be assessed in terms of their core methodologies, advantages, and constraints, all of which have significantly shaped current QE practices.

By 2013, with the introduction of word embedding like Word2Vec and GloVe, QE benefited from context-independent word representations that improved translations' semantic understanding. However, despite their progress, these models still failed to capture the intricacies of translation context and multilingual diversity fully.

The QuEst model (Specia *et al.*, 2020), marked a significant shift toward using traditional machine learning for QE, followed by its refinement into QuEst++. These models improved performance by incorporating additional features but were still limited in their flexibility and effectiveness across different language pairs. QUETCH and NuQE Predictor-Estimator (Kepler *et al.*, 2019) were among the first to experiment with neural network-based models.

MARMOT and POSTECH introduced neural-based QE models. MARMOT used Conditional Random Fields for word-level QE, focusing on feature-based modelling, while POSTECH introduced the Predictor-Estimator architecture, which removed the need for manual feature engineering. Though POSTECH marked a leap forward by incorporating neural networks, it required substantial pre-training and was limited by its computational demands. These models laid the groundwork for the shift toward more flexible neural architectures in the following years. Figure 2 explains developmental stages of each phase of LLM Models.

The advent of transformer-based models, particularly BERT(Devlin *et al.*, 2018), revolutionized QE. BERT's bidirectional understanding of context provided a deeper insight into translation quality, though its resource-intensiveness remained a challenge for widespread use. The release of multilingual transformers, such as MBERT (Devlin *et al.*, 2018) sought to extend these capabilities across languages. However, Mbert's inconsistent performance across languages underscored the limitations of multilingual models trained on diverse data.

Developing translation-specific models, such as TLM, and models like XLM enhanced cross-lingual performance, significantly improving the handling of multilingual tasks. DeepQuest (Alva-Manchego *et al.*, 2021a) and OpenKiwi (Kepler *et al.*, 2019) played a pivotal role by providing an open-source framework for QE, allowing greater experimentation and comparison across architectures. Additionally, multilingual models like mDistilBERT and XLM-RoBERTa (Conneau *et al.*, 2020) improved upon previous transformers by offering better cross-lingual abilities. This period also saw the development of TransQuest (Ranasinghe, Orăsan and Mitkov, 2021)a model that combined XLM-R embeddings for better performance on sentence-level and word-level QE tasks. The introduction of COMET(Qian *et al.*, 2024) a reference-less evaluation model, marked another breakthrough, with significant improvement in multilingual contexts, thanks to its ability to assess translations without human references.

Various methods have shown promising results in the QE shared task at WMT (Alva-Manchego *et al.*, 2021b; Heafield, Zhu and Grundkiewicz, 2021b; Bhattacharyya *et al.*, 2023b)though most rely on supervision and training (Kanojia *et al.*, 2021; Deoghare, Kanojia and Ranasinghe, 2023) With the surge of LLMs, their application in translation quality assessment has gained traction, as seen in GEMBA, a zero-shot prompting approach

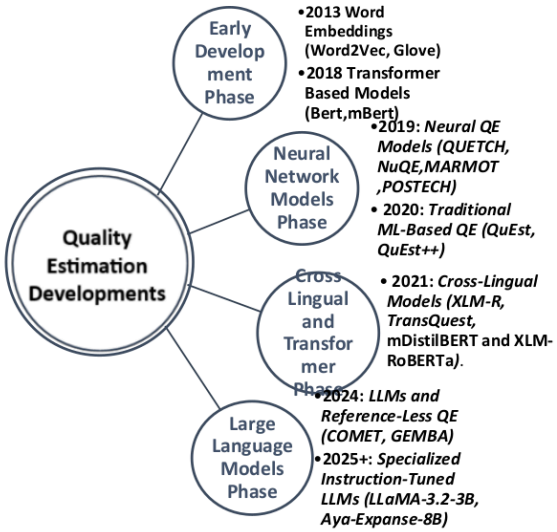


Figure 2: Detailed Evolution and Developmental Stages LLM's Models.

For Direct Assessment score prediction using GPT-4 (Fernandes *et al.*, 2023; Kocmi and Federmann, 2023).

Models like LlaMA-2 (Iyer *et al.*, 2024) expanded the scope by integrating monolingual and parallel data, enhancing their ability to deal with diverse languages and domains. Further advancements have focused on the specialization of LLMs for specific tasks through techniques such as instruction tuning and adversarial evaluation. These improvements offer more efficient and adaptable models for various domains and translation contexts. The proposed approach improves LlaMA-3.2-3B-Instruct, LlaMA-3.2-3B-Instruct, Mistral-7B-Instruct-v0.2, and Gemma-1.1-4B-IT, which are LLMs used in natural language processing tasks due to their unique strengths and capabilities.

2.2 RESEARCH GAP

Despite significant progress in the development of Machine Translation Quality Estimation (MT QE) systems, several important research challenges remain, particularly in the context of low-resource language pairs such as English–Hindi (en–hi) and English–Marathi (en–mr). Additionally, the integration of zero-shot and few-shot prompting techniques using Large Language Models (LLMs) for QE has not been fully explored in current literature.

First, existing QE models—such as TransQuest, COMET—exhibit a strong dependence on human-annotated direct assessment (DA) scores. While effective in controlled settings, this reliance on costly and time-consuming human annotation significantly limits the scalability of these models to under-resourced languages and emerging domains. This constraint poses a critical barrier to developing adaptable QE systems for diverse translation contexts, including those involving Indic language pairs like en–hi and en–mr.

Second, although multilingual transformer-based models (e.g., Mbert, XLM-R) have demonstrated success in cross-lingual transfer, their performance remains inconsistent across languages. The so-called “curse of multilingualism” continues to hinder the generalizability of these models, especially when applied to languages with limited training data. Consequently, current QE systems struggle to maintain robust performance across linguistically diverse translation tasks.

Third, existing QE frameworks lack advanced semantic reasoning capabilities necessary for handling complex translation phenomena such as idiomatic expressions, domain-specific terminology, and ambiguous constructions. While contextual models like BERT and XLM improve token-level representation, they often fall short in disambiguating nuanced meanings.

Fourth, the dominant use of traditional lexical-similarity metrics (e.g., BLEU, TER, chrF) for translation quality assessment overlooks semantic adequacy and fluency. Although recent metrics like COMET offer reference-free alternatives, their efficacy in zero-shot QE scenarios particularly for low-resource languages remains insufficiently validated.

Finally, while instruction-tuned LLMs (e.g., LlaMA-2, Mistral, Gemma) have demonstrated remarkable capabilities in various NLP tasks, their potential for instruction-based QE remains under-investigated. Most prior work focuses on generation and reasoning tasks, rather than using LLMs for evaluating translation quality, especially under few-shot and zero-shot conditions.

This study integrates zero-shot and few-shot prompting techniques by utilizing instruction-tuned large language models (LLMs), such as LlaMA-3.2-3B-Instruct, Mistral-7B-Instruct-v0.2, and Gemma-1.1-4B-IT. This study focuses on multilingual fine-tuning specifically for low-resource language pairs, namely English–Hindi (en–hi) and English–Marathi (en–mr), using WMT datasets. This approach reduces dependence on human-labeled direct assessment (DA) scores and promotes broader generalizability in multilingual quality estimation.

This research introduces a novel instruction-based quality estimation framework that leverages prompt-based, instruction-following inference to move beyond traditional lexical matching metrics and enable a more semantically informed evaluation of translation quality. Lastly, the performance of the fine-tuned LLMs is comprehensively benchmarked against established QE models such as TransQuest and COMET.

2.3 Theoretical Framework

This research is grounded in the theoretical domains of Machine Translation (MT) Quality Estimation (QE), multilingual Natural Language Processing (NLP), and the application of Large Language Models (LLMs). The study synthesizes linguistic principles with state-of-the-

art deep learning methodologies to enhance QE across low-resource language pairs, specifically English–Hindi (en–hi) and English–Marathi (en–mr). Drawing upon advances in transfer learning, zero-shot learning (ZSL), Few-Shot Learning (FSL) prompting, and instruction tuning, the proposed framework aims to overcome limitations in current QE practices by exploiting the reasoning capabilities of instruction-tuned LLMs.

2.3.1 Quality Estimation (QE) Theory

Quality Estimation has undergone a significant transformation—from early feature-engineered models such as QuEst and QuEst++ to neural architectures like MARMOT and POSTECH, and later to transformer-based frameworks including TransQuest, COMET. Despite these advances, most systems continue to rely on direct assessment (DA) scores from human-annotated data, limiting scalability across new domains and under-resourced languages.

QE is fundamentally the task of evaluating the quality of machine-generated translations without access to human references. Traditional evaluation metrics such as BLEU, TER, and BERTScore compare system outputs to reference translations, but fail in zero-reference scenarios—particularly prevalent in low-resource settings like en–hi and en–mr. This study investigates how instruction-fine-tuned LLMs like LlaMA-3.2-3B-Instruct, Mistral-7B-Instruct-v0.2, Gemma-1.1-4B-IT, can conduct QE with minimal or no supervision, leveraging their pre-trained multilingual and contextual reasoning abilities to predict translation quality directly from source and target sentence pairs.

2.3.2 Multilingual NLP and Transfer Learning

Multilingual NLP has evolved with the introduction of models such as Mbert, XLM, and XLM-R, which enable zero-shot and Few-shot transfer across languages. However, these models often exhibit inconsistent performance on low-resource pairs due to the "curse of multilingualism," were increased language support compromises model accuracy.

This research addresses these issues by fine-tuning LlaMA-3, Mistral, and Gemma on MT QE datasets involving en–hi and en–mr. The objective is to improve QE generalization without over-reliance on human-annotated data. Cross-lingual embeddings are used to capture deeper semantic relations between source and target texts, while instruction tuning is employed to align the models' outputs with task-specific quality criteria. These enhancements are expected to boost QE accuracy, especially for low-resource translation scenarios.

2.3.3 Zero-Shot and Few-Shot Learning

Zero-shot learning (ZSL) enables models to generalize to unseen language pairs without explicit retraining, a crucial capability for MT QE in underrepresented languages like en–hi and en–mr. Similarly, few-shot learning allows models to learn QE strategies from a minimal number of annotated examples, making evaluation feasible in data-scarce environments.

This study incorporates ZSL and few-shot prompting paradigms using instruction-tuned LLMs to perform quality estimation without human references.

2.3.4 Instruction Tuning in Large Language Models for QE

The emergence of LLMs such as LlaMA-2, Mistral, and Gemma represents a paradigm shift in QE research. Unlike traditional supervised QE models, these LLMs can be instruction-tuned or prompted to execute evaluation tasks with minimal supervision. This flexibility enables the development of adaptable QE systems that do not rely on extensive domain-specific annotations.

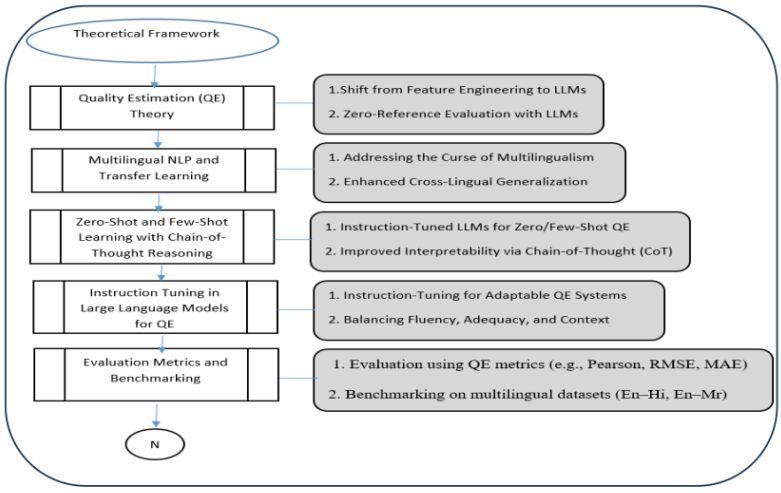
In this study, instruction tuning is employed to teach models task-specific behaviour for QE, while few-shot prompting provides minimal examples to guide their inferences. These techniques are evaluated to determine whether instruction-based approaches can rival or outperform conventional fine-tuned QE models in terms of accuracy and interpretability. Special focus is placed on how LLMs manage to balance fluency, semantic adequacy, and contextual correctness, particularly in zero-reference settings across en–hi and en–mr language pairs.

2.3.5 Evaluation Metrics and Benchmarking

The performance of the proposed LLM-based QE framework is assessed using both traditional and modern evaluation metrics. These include Direct Assessment (DA) scores and lexical similarity metrics such as BLEU, TER, and BERTScore, alongside newer context-aware scores like COMET. The framework's results are benchmarked against state-of-the-art QE systems including TransQuest, COMET

To evaluate model robustness across different data availability scenarios, experiments are conducted under both zero-shot and few-shot conditions. The goal is to validate whether LLMs can effectively generalize to new language pairs and domains while delivering competitive or superior performance relative to existing models. This benchmarking informs the broader

objective of creating scalable, instruction-driven QE systems that perform reliably across multilingual contexts, especially for low-resource combinations like en–hi and en–mr.



Flowchart 1: - Theoretical Framework

2.4 Historical Context

The evolution of Machine Translation (MT) Quality Estimation (QE) has mirrored the broader trajectory of natural language processing, moving from feature-based methods to transformer-based and instruction-tuned paradigms. Early QE systems, such as QuEst and QuEst++, relied on statistical learning models with manually engineered linguistic features. These models offered limited scalability and language adaptability. The introduction of neural architectures, including MARMOT and POSTECH, brought predictor-estimator frameworks into focus, allowing for automatic feature learning but still dependent on supervised training data.

The advent of transformer-based models like BERT and XLM marked a pivotal shift, enabling deep contextual representation and improved cross-lingual transfer. This led to the development of multilingual transformers such as XLM-R and mDistilBERT, which expanded QE capabilities to multiple languages, albeit with performance degradation on low-resource pairs—commonly referred to as the "curse of multilingualism."

Frameworks such as OpenKiwi and deepQuest democratized access to QE research, while models like TransQuest and COMET introduced robust sentence-level evaluation by

leveraging deep semantic embeddings and reference-free architectures. Despite these advances, these models still heavily depend on human-annotated Direct Assessment (DA) scores, limiting scalability to unseen language pairs or novel domains.

The recent rise of Large Language Models (LLMs) such as LlaMA-2, Mistral, and Gemma has redefined QE paradigms through instruction tuning and prompt-based inference. These models enable zero-shot and few-shot learning strategies, allowing for quality estimation without explicit retraining on annotated datasets. Additionally, the incorporation of Chain of Thought (CoT) reasoning has enhanced the interpretability and linguistic robustness of LLM-driven QE outputs.

In particular, the application of these LLMs to low-resource language pairs, such as English–Hindi (en–hi) and English–Marathi (en–mr), signifies a crucial advancement in multilingual QE. By reducing dependency on labelled resources and enabling contextualized evaluation through instruction-based learning, this research contributes to a more scalable, explainable, and language-agnostic QE ecosystem.

2.5 METHODOLOGICAL REVIEW

The field of Machine Translation (MT) Quality Estimation (QE) has experienced a profound methodological evolution—from early rule-based systems to deep learning architectures and, more recently, instruction-tuned Large Language Models (LLMs). This section critically reviews these paradigms, highlighting their strengths, limitations, and future potential, with a particular focus on their applicability to low-resource language pairs such as English–Hindi (en–hi) and English–Marathi (en–mr).

2.5.1 Early Rule-Based and Statistical Approaches

Initial QE systems were grounded in rule-based and statistical learning techniques, relying on manually engineered linguistic features to assess translation quality. Notable examples include QuEst and QuEst++, which extracted syntactic and semantic cues from source and translated texts. While these systems demonstrated basic linguistic insight, they suffered from limited scalability and cross-lingual adaptability. Their performance was heavily contingent on human expertise in feature design, making them ill-suited for rapid deployment across diverse or low-resource language pairs.

2.5.2 Neural Network-Based Architectures and Predictor-Estimator Models

The introduction of neural network-based architectures marked a turning point in QE methodology. Models such as QUETCH and the NuQE predictor-estimator framework replaced handcrafted features with data-driven representations learned through neural layers. Systems like MARMOT and POSTECH incorporated Conditional Random Fields (CRFs) and multi-stage predictor-estimator pipelines to improve sentence and word-level scoring accuracy. Despite these advances, such models remained reliant on substantial quantities of annotated data, which posed challenges for low-resource languages. Moreover, their fixed architectures lacked adaptability across domains and translation contexts.

2.5.3 Transformer-Based Models and the Emergence of Contextual Representations

Transformer-based models brought further refinement, particularly with the introduction of BERT which captured deep bidirectional context through self-attention mechanisms. Multilingual variants like Mbert and XLM enabled cross-lingual knowledge transfer, although with uneven performance on typologically diverse or underrepresented languages. This motivated the development of models like XLM-R and TLM, which provided better multilingual representation learning.

Open-source platforms like OpenKiwi and deepQuest accelerated research in this area by offering modular tools for benchmarking and experimentation. TransQuest built upon XLM-R embeddings to deliver state-of-the-art results on both sentence-level and word-level QE tasks. However, these transformer-based systems were computationally intensive, posing practical limitations in environments with limited processing capabilities.

2.5.4 Reference-Less QE and Advancements in Large Language Models (LLMs)

Recent methodological innovations have focused on reducing dependency on human-annotated references. COMET introduced reference-less evaluation strategies that assess translation quality solely based on source and hypothesis, enabling greater flexibility in real-world MT workflows.

Simultaneously, the emergence of instruction-tuned LLMs such as LlaMA-3, Mistral, and Google's Gemma has redefined QE by enabling zero-shot and few-shot estimation via prompting. These models leverage vast pretraining across languages and domains, allowing for effective quality estimation without the need for retraining on annotated QE datasets. Notably, prompting strategies like Chain of Thought (CoT) reasoning have enhanced transparency by decomposing the estimation process into interpretable steps.

LLMs are particularly valuable in low-resource settings such as en–hi and en–mr, where labelled QE datasets are scarce. By combining general linguistic reasoning with contextual evaluation, these models offer a scalable and adaptable solution to multilingual QE.

2.5.5 Challenges in Existing Methodologies

Despite substantial progress, several methodological challenges remain:

1. **Data Dependence**: Supervised QE models continue to rely heavily on large annotated corpora, which are costly and time-intensive to develop.
2. **Resource Constraints**: Transformer-based and LLM-driven methods require high computational overhead, limiting feasibility in resource-constrained environments.
3. **Multilingual Performance Gaps**: Even advanced models like XLM-R and COMET exhibit variable performance across different language pairs, with notable degradation for low-resource languages.
4. **Lack of Interpretability**: Many deep learning models operate as opaque black boxes; although techniques like CoT reasoning aid interpretability, more intrinsic explanations are needed.
5. **Domain Adaptation**: Existing QE models often underperform when applied to domain-specific texts (e.g., legal, medical), where terminology and structure diverge from general corpora.

2.5.6 Future Methodological Directions

To address these limitations, future QE methodologies should prioritize the following directions:

- **Lightweight and Efficient Architectures**: Develop QE models that offer strong performance with lower computational requirements, enabling broader accessibility.
- **Hybrid Paradigms**: Combine fine-tuned transformers with prompt-based LLM inference to leverage both precision and adaptability.
- **Enhanced Multilingual Generalization**: Train on linguistically diverse datasets to improve robustness across underrepresented language pairs.
- **Self-Explainable QE Systems**: Design architectures that natively justify their quality judgments, improving transparency and trustworthiness.

- **Adaptive Learning Mechanisms**: Implement QE models that can incrementally learn from user feedback or in-context examples, supporting dynamic refinement over time.

Figure 5 explains the detailed Methodological Review ,which explains each stage clearly.

Flowchart 2 – Methodological Review

2.6 SUMMARY

This literature review explores the evolution and current landscape of Machine Translation Quality Estimation (MT QE), emphasizing the transition from traditional methods to advanced

neural and transformer-based models. Early QE approaches utilized context-independent word embeddings like Word2Vec and GloVe, offering limited semantic understanding. Traditional machine learning models, such as QuEst and its enhancement QuEst++, laid foundational work by introducing hand-crafted features. Neural models like QUETCH, NuQE, MARMOT, and POSTECH improved flexibility and reduced the need for manual feature engineering, though they required significant computational resources and pretraining.

A pivotal shift occurred with the emergence of transformer-based models like BERT, mBERT, and XLM, which improved context comprehension and multilingual performance. However, inconsistencies remained in low-resource language scenarios. Tools like deepQuest, OpenKiwi, and TransQuest provided open-source frameworks to test and compare models, while newer multilingual transformers like XLM-RoBERTa and mDistilBERT further refined cross-lingual capabilities. COMET introduced a reference-less evaluation approach, enhancing scalability across languages. The rise of instruction-tuned Large Language Models (LLMs) such as LLaMA-2, GPT-4, Mistral, and Gemma brought zero-shot and few-shot learning into QE tasks.

Despite these advancements, significant gaps persist. QE systems remain heavily dependent on human-labelled DA scores and struggle with low-resource languages like English–Hindi (en–hi) and English–Marathi (en–mr). Existing models lack semantic reasoning needed for idiomatic expressions and ambiguous constructions. Moreover, traditional metrics like BLEU and TER fail to assess fluency and adequacy in zero-reference settings.

This research addresses these limitations by integrating instruction-based prompting with fine-tuned LLMs (LLaMA-3.2-3B, Mistral-7B, Gemma-1.1-4B-IT) using Few-shot and Zero-Shot reasoning. It aims to reduce reliance on reference translations and DA scores while enhancing semantic understanding. The framework benchmarks these models against traditional QE systems, validating their efficacy across multilingual, low-resource contexts using both automatic and LLM-generated quality scores.

CHAPTER-3

RESEARCH METHODOLOGY

3.1 INTRODUCTION

This research aims to evaluate the effectiveness of Large Language Models (LLMs) in Machine Translation (MT) Quality Estimation (QE), with a particular focus on multilingual settings involving low-resource languages such as Hindi and Marathi. The goal of the methodology is to systematically compare the performance of several state-of-the-art LLMs across different translation quality tasks, determining their accuracy, scalability, and suitability for real-world applications.

To achieve this, we have adopted a quantitative, comparative experimental approach. This design enables objective performance measurement and statistical comparison across multiple models and translation tasks. A quantitative framework is especially appropriate in this context, as it allows for consistent evaluation using established metrics such as correlation scores and quality estimation benchmarks. The comparative nature of the design supports direct model-to-model assessment, while the experimental setup ensures that controlled variables are used across all tests to maintain fairness and reproducibility.

We focused on LLMs in modern natural language processing pipelines and their proven potential in both zero-shot and fine-tuned configurations for translation-related tasks. LLMs offer contextual reasoning, multilingual adaptability, and scalability, making them promising candidates for QE, particularly in scenarios where traditional reference-based metrics are unavailable or impractical. Furthermore, the inclusion of multilingual QE tasks addresses the pressing need for robust evaluation in underrepresented language pairs, where annotated data is scarce.

The methodology leverages several high-quality benchmark datasets to support this investigation. These include the WMT23 dataset (sourced from Hugging Face), which provides standard QE evaluation data, and two specific test datasets: QE-2023-EnHi-Test and QE-2023-EnMr-Test, which are tailored for English-to-Hindi and English-to-Marathi translation quality assessment, respectively. These datasets encompass source sentences, MT outputs, and in some cases, human-generated references or DA scores.

Model performance is evaluated across a spectrum of tasks, including sentence-level QE, summarization, and generative tasks, using a mix of open-source models—such as LLaMA-3.2-3B-Instruct, Mistral-7B-Instruct-v0.2, and Gemma-1.1-4B-IT. Evaluations are conducted

both on pre-trained and fine-tuned model variants to analyze the benefits of instruction tuning and multilingual training.

3.2 RESEARCH METHODOLOGY

3.2.1 DATA SELECTION

In this research, the evaluation of Machine Translation Quality Estimation (MT QE) is grounded in high-quality, human-annotated data derived from the WMT23 Quality Estimation Shared Task. This dataset is specifically curated to support benchmarking in sentence-level and word-level QE tasks across multiple language pairs. The present study focuses on the low-resource language pairs of English–Hindi and English–Marathi, which represent linguistically and morphologically rich target languages with limited labelled data in the context of translation quality estimation.

The WMT23 QE dataset consists of parallel corpora that pair a source sentence (in English) with its machine-translated counterpart (in Hindi or Marathi). Each pair is annotated with quality labels at sentence level as Sentence-Level Annotations Each source–translation pair is assigned a Direct Assessment (DA) score. This score, ranging from 0 to 100, reflects the degree of adequacy and fluency as perceived by human annotators. These scores form the basis for regression-based QE tasks, where the objective is to predict a continuous value indicating translation quality.

Both Hindi and Marathi are Indo-Aryan languages written in Devanagari script, but they exhibit syntactic and semantic distinctions that challenge general-purpose language models. The dataset exhibits the following properties:

Selected Datasets:

1. **WMT23 QE Dataset (via Hugging Face):**
   This dataset includes source sentences, machine-generated translations, and reference translations across multiple language pairs. It also provides human-annotated quality scores, enabling correlation-based performance evaluation. It serves as a foundational benchmark for sentence-level and document-level QE tasks.
2. **QE-2023-EnHi-Test:**
   This dataset focuses specifically on English-to-Hindi translation quality estimation. It consists of parallel source-translation pairs accompanied by reference scores or metadata,

providing a targeted dataset for evaluating model performance in low-resource language pairs.

3. **QE-2023-EnMr-Test:**
   Designed for English-to-Marathi QE evaluation, this dataset complements the En-Hi data by expanding the language coverage. It enables a comparative study of model performance across two linguistically distinct Indian languages.

The dataset used in this study sourced from the WMT23 Quality Estimation (QE) shared task was provided in two pre-defined subsets: train.tsv and dev.tsv for both en-hi and en-mr. These files contain sentence-level Direct Assessment (DA) scores for machine-translated outputs in English–Hindi and English–Marathi, which are annotated by human evaluators.

- **train.tsv**: This file contains the main training corpus and was used for training tasks.
- **dev.tsv:** This file served as the evaluation/test set. No tuning or prompt adaptation was conducted on this subset to preserve fairness and to measure zero-shot generalization of LLMs.

Both files were structured with consistent columns including:

- **source:** The original English sentence
- **mt:** The machine-translated output (Hindi or Marathi)
- **score:** The DA-based QE score in the normalized range [0, 100]

Table 1: Summary Statistics of train.tsv and dev.tsv[en-hi]

Metric	**train.tsv**	**dev.tsv**
Number of Sentence Pairs	7,000	1,000
Language Pairs	En–Hi	En–Hi
Mean QE Score	80.825	80.762
Standard Deviation	7.942	7.982
Min–Max Score Range	14.25 -98.25	31.25-98.5
Missing Values	0	0

Table 2: Summary Statistics of train.tsv and dev.tsv[en-mr]

Metric	train.tsv	dev.tsv
Number of Sentence Pairs	26,000	1,000
Language Pairs	En–Mr	En–Mr
Mean QE Score	70.077	69.805
Standard Deviation	10.155	10.949
Min–Max Score Range	0.5-176.5	2.5-168.5
Missing Values	**0**	**0**

As shown in Tables 1 and 2, the English–Hindi dataset exhibited a relatively narrow QE score range with a high mean around 80, indicating overall good translation quality. In contrast, the English–Marathi dataset demonstrated a broader score distribution, with scores ranging from 0.5 to 176.5, and a lower mean, suggesting greater variability and potentially noisier or more diverse translations. No missing values were observed in any of the datasets.

3.2.2 DATA PREPROCESSING

In this study, we leverage instruction-tuned large language models, each of which relies on a specific tokenizer as part of its preprocessing pipeline. Tokenization is a crucial step that converts raw input text into model-readable token IDs, maintaining consistency with the model's original pretraining setup.

3.2.2.1. TransQuest Tokenizer

TransQuest uses sentence-transformer architectures based on models like XLM-RoBERTa or DistilBERT. TransQuest relies on Hugging Face's AutoTokenizer associated with its transformer backbone (typically xlm-roberta-base). The input is a pair of sentences—the source (original) and the hypothesis (machine-translated output). These are tokenized as sentence pairs with special separators (<s>, </s>).

3.2.2.2 COMET Tokenizer

COMET uses the SentencePiece tokenizer from its transformer backbone (e.g., xlm-roberta-large). COMET typically processes triplets consisting of the source sentence, the MT hypothesis, and the reference translation (for reference-based scoring) or just source and hypothesis (for QE). Sentences are tokenized and concatenated using separators, often with special tokens: [CLS] source </s> hypothesis </s> reference.

3.2.2.3 LLaMa Tokenizer

LLaMA-3.2-3B-Instruct employs Meta's SentencePiece-based tokenizer with a vocabulary size of approximately 128K. It operates at the byte level and supports multilingual UTF-8 input. The tokenizer segments input into subword units while preserving semantic integrity, and formats instruction-style prompts using special role tokens like <|start_header_id|>user<|end_header_id|>

3.2.2.4 Mistral Tokenizer

Mistral-7B-Instruct uses a Byte-Pair Encoding (BPE) tokenizer aligned with Hugging Face's tokenizers library. The tokenizer has a vocabulary of around 32K tokens and works at the byte level, allowing effective processing of diverse linguistic inputs. It is optimized for dialogue-style input using role-indicative tokens such as <|user|> and <|assistant|>.

3.2.2.5 Gemma Tokenizer

Gemma-1.1-4B-IT utilizes Google's SentencePiece tokenizer with a 32K token vocabulary. Designed for instruction-following tasks, it processes text using a ChatML-like structure with markers such as <start_of_turn>user<end_of_turn>. Its byte-level encoding supports consistent subword tokenization across varied sentence structures.

3.2.3 EVALUATION

This section presents a comprehensive analysis of the proposed QE framework's performance, integrating both quantitative metrics and qualitative insights to assess its reliability, efficiency, and generalization capabilities.

The evaluation was conducted using established QE metrics. Pearson and Spearman correlation coefficients were used to assess the relationship between predicted QE scores and human Direct

Assessment (DA) scores—capturing both the consistency of absolute values and the agreement in ranking order. In addition, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were employed to quantify prediction errors, where lower values indicate better alignment with human judgment. To evaluate computational practicality, inference time and GPU memory consumption were measured, allowing an assessment of the resource efficiency of LLaMA-3, Mistral, and Gemma within real-world MT workflows.

On the qualitative side, the evaluation focused on the contextual reasoning abilities of the large language models. Justification quality was analyzed by examining the interpretability of model-generated feedback accompanying QE scores. More informative and contextually grounded explanations were found to improve user trust in the system. Error behavior analysis showed that the models generally performed well on fluent translations, but struggled with syntactically or semantically degraded outputs—particularly in the low-resource English–Marathi (En–Mr) setting. Despite this, the models demonstrated cross-lingual generalization, performing reasonably well on both Hindi and Marathi translation pairs, although results varied depending on linguistic complexity and training exposure.

A comparative benchmarking was carried out across LLaMA-3, Mistral, and Gemma on identical test sets to ensure a fair evaluation. The comparison focused on prediction accuracy, explanation clarity, and computational efficiency. The findings suggest that LLM-based QE models can rival or even surpass traditional QE techniques in terms of correlation with human judgment. However, optimal performance depends on effective prompt engineering, model selection, and an evaluation strategy that balances prediction quality, interpretability, and resource constraints.

3.3 PROPOSED METHOD

The proposed system harnesses the capabilities of instruction-tuned large language models (LLMs) to perform automatic quality estimation (QE) of machine-generated translations. The proposed solution integrates both encoder-based and decoder-based models under a unified framework for Translation Quality Estimation:

- **Encoder-only models** (TransQuest, COMET) are used for efficient, regression-based scoring, leveraging pretrained multilingual encoders with minimal overhead.

- **Decoder-only models** (LLaMA, Mistral, Gemma) are used for instruction-based qualitative analysis and score generation, providing deeper insight into translation errors and contextual alignment.

  Let us Understand historically,

3.3.1 Transformer Architecture: The Foundational Backbone

The Transformer architecture, first introduced by (Vaswani *et al.*, 2017), has become the dominant paradigm in natural language processing due to its scalability, parallelism, and capacity to model long-range dependencies. The architecture consists of two main components:

- **Encoder**: Processes input sequences through self-attention and feed-forward layers to generate context-rich representations.
- **Decoder**: Generates outputs by attending to both the encoder's representations and previously generated tokens through masked self-attention.

The core mechanism multi-head self-attention allows the model to weigh the relevance of every word in a sequence with respect to all others, providing an effective means for contextual understanding.

This architecture serves as the basis for both encoder-only and decoder-only variants, which are adapted to different types of tasks including classification, regression, and generative modeling.

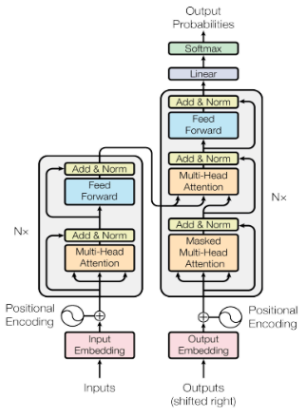


Figure 3: Transformer Architecture (adapted from Vaswani et al., 2017)

The Transformer follows this overall architecture using stacked self-attention and point-wise, fully connected layers for both the encoder and decoder, shown in the left and right halves of Figure 3, respectively.

3.3.2 Encoder-Only Architectures for Quality Estimation

Encoder-only models, such as BERT, RoBERTa, and XLM-RoBERTa, use the encoder stack of the Transformer to encode input text into dense vector representations. These models are well-suited for discriminative tasks like sentence classification, semantic similarity, and regression-based quality estimation, where a complete input is mapped to a scalar output or class label.

a. TransQuest

- Built upon multilingual encoder models like XLM-RoBERTa, TransQuest represents sentence pairs using pooled embeddings and computes a continuous quality score using a regression head. It concatenates the source sentence and its translation into a single input sequence using special tokens like [SEP].
- Trained on human-annotated direct assessment (DA) scores, TransQuest learns to predict the adequacy and fluency of a translation without the need for reference translations. Fast inference, fewer parameters, and language-agnostic performance make it a practical choice for low-resource or real-time settings.

b. COMET (Crosslingual Optimized Metric for Evaluation of Translation)

- COMET builds on XLM-R or InfoXLM, which are large-scale multilingual encoder models. It encodes triplets: source, hypothesis, and optionally reference translations. The model is trained using regression loss on human-annotated scores and optimizes for high correlation with human judgment.
- COMET models are capable of generalizing to unseen language pairs thanks to pretraining on multilingual corpora. Unlike traditional n-gram or lexical overlap-based metrics (e.g., BLEU), COMET captures semantic similarity and contextual fidelity, making it ideal for nuanced QE tasks.

Encoder-only QE models are particularly appealing in industrial settings due to their fast inference, minimal prompt engineering, and robust correlation with human evaluation metrics.

3.3.3 Decoder-Only Architectures for Quality Estimation

Decoder-only models are built from the decoder component of the Transformer and are designed primarily for autoregressive language modeling. These models are trained to predict the next token in a sequence and can be adapted for downstream tasks through prompt engineering and instruction tuning. In the context of QE, decoder-only models are employed by framing the problem as a text generation task, where the model is prompted with the source and hypothesis, and asked to generate either a quality score or a qualitative justification.

a. LLaMA-3.2-3B-Instruct

LLaMA-3 is a decoder-only causal language model designed for high-efficiency instruction-following tasks. The model is prompted with structured instructions such as: "*Evaluate the quality of the following translation and provide a score out of 100:*" allowing it to generate both a score and a rationale.

LLaMA-3 shows robust reasoning abilities and zero-shot performance across diverse languages, despite not being fine-tuned explicitly for QE. The autoregressive decoding makes inference slower than encoder-only models, and performance is highly dependent on prompt quality.

b. Mistral-7B-Instruct

Mistral is a dense decoder-only model optimized for high-throughput inference and multilingual instruction-following. Mistral can be prompted similarly to LLaMA and exhibits strong performance in generating interpretable justifications alongside quality scores. Mistral supports long-context reasoning and is known for its computational efficiency, making it suitable for deployment in real-time QE pipelines.

c. Gemma-1.1-4B-IT

Gemma is a decoder-only instruction-tuned model developed by Google, fine-tuned for safety, alignment, and multi-turn dialogue. With minimal prompt adaptation, Gemma can evaluate translation quality and output human-like justifications or ratings.

Demonstrates competitive performance on multilingual tasks, especially when dealing with complex linguistic phenomena like ambiguity and idiomatic usage.

Decoder-only models enable rich qualitative insights by generating textual feedback and scoring simultaneously, enhancing the interpretability and trustworthiness of QE systems.

To address the challenge of machine translation quality estimation (QE), the proposed approach reframes the problem to better align with the capabilities of instruction-tuned large language models (LLMs). Specifically, the method formulates QE as a classification task using prompt-based inference, thus improving interpretability, scalability, and integration with downstream systems. The key steps of the proposed methodology are described below:

Step1. Problem Formulation:

Although QE is typically a regression problem (with scores between 0 and 1), we transform it into a classification problem by defining discrete quality classes:

- Class 0 – Poor translation (QE score: 0.0–0.3)
- Class 1 – Moderate translation (QE score: 0.3–0.7)
- Class 2 – High-quality translation (QE score: 0.7–1.0)

This simplifies model output and is more interpretable for downstream applications such as automatic filtering or reranking.

Step2. Input Formatting:

Each sample is converted into a structured prompt suitable for LLM-based classification:

Source: <src> The original English sentence </src>
Translation: <tgt> The machine-generated translation (Hindi/Marathi) </tgt>
Task: Classify the quality of the translation into one of the following:
0 (Poor), 1 (Moderate), or 2 (Excellent). Justify your answer briefly.

Step3. Model Architecture:

Open-source LLMs are used, including:

- LLaMA-3–3B-Instruct
- Mistral-7B-Instruct
- Gemma-1.1–4B-IT

These models are prompted in zero-shot or few-shot settings to generate both a class label and a justification.

Step4. Label Transformation:

Human Direct Assessment (DA) scores were mapped to the appropriate class labels (0, 1, or 2) during preprocessing. This ensures alignment between numeric scores and categorical outputs.

Step5. Training and Inference

- **For fine-tuning:** The model is trained on labelled (class) data using supervised learning with cross-entropy loss.
- **For inference:** The model predicts a class label based on its instruction-following ability, optionally with confidence estimation.

Step6. Output

The model generates:

- A predicted class label (0, 1, or 2)
- A textual explanation supporting the classification decision
- Optional: confidence scores or probability distribution over classes

This classification-oriented, instruction-driven framework enhances model transparency, supports multilingual quality estimation (e.g., English–Hindi, English–Marathi), and simplifies integration into machine translation pipelines by offering both quantitative scores and qualitative insights.

3.4 TOOLS

This section outlines the tools, libraries, and technologies utilized for the development and evaluation of the Machine Translation Quality Estimation (MT QE) system.

3.4.1 Programming Environment

- **Python 3.10+**
  Python was the primary programming language used throughout the project for data handling, model interfacing, evaluation, and visualization tasks due to its simplicity and rich ecosystem of ML/NLP libraries.
- **Google Collaboratory (Colab)**
  Google Colab served as the primary development environment. It provides free access to GPU and TPU resources, enabling faster execution of large-scale models. Specifically, the NVIDIA Tesla T4 GPU was utilized for inference and fine-tuning tasks.

3.4.2 Libraries and Frameworks

- **NumPy**

  Employed for efficient numerical operations, especially for processing tensor outputs and computing evaluation metrics.

- **Pandas**

  Used for structured data manipulation, especially for reading and processing .tsv files that contain source sentences, machine-translated sentences, and quality scores.

- **Scikit-learn**

  Provided implementations for key evaluation metrics such as Pearson correlation, Spearman correlation, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE).

- **Plotly**

  Utilized for creating interactive and publication-quality visualizations to represent the distribution of scores, correlation comparisons, and model behaviour.

3.4.3 NLP and Deep Learning Libraries

- **Hugging Face Transformers**

  A core framework used for loading and interacting with state-of-the-art language models such as:

  - meta-llama/Llama-3-3B-Instruct
  - mistralai/Mistral-7B-Instruct-v0.2
  - google/gemma-1.1-4b-it

  These models were integrated using prompt-based approaches for sentence-level quality estimation without explicit training. Some other Models like TransQuest, Comet also created for comparative analysis.

- **SentenceTransformers**

  Employed for baseline QE experiments and feature extraction using pre-trained multilingual sentence embedding models like:

  - paraphrase-multilingual-mpnet-base-v2
  - LaBSE
  - all-mpnet-base-v2

3.4.4 Prompt Engineering

- Carefully designed prompts were constructed to guide large language models (LLMs) to predict sentence-level translation quality. Prompts were iteratively refined and adapted to suit the language pair and the instruction-following capabilities of each model.

3.5 SUMMARY

This research employs a comprehensive and structured methodology to investigate the effectiveness of instruction-tuned Large Language Models (LLMs) for the task of Machine Translation Quality Estimation (MT QE), particularly for low-resource language pairs such as English–Hindi and English–Marathi. The methodology is designed to support both regression and classification paradigms, with a strong emphasis on prompt-based inference, minimal supervision, and evaluation against human-annotated benchmarks.

The methodology consists of the following key components:

Data Preprocessing: The source datasets (WMT23 QE) were cleaned, normalized, and aligned to ensure high-quality sentence pairs. Language-specific processing, such as Unicode normalization and token handling, was applied, followed by dataset splitting and format conversion suitable for LLM prompting.

Model Selection: A diverse set of open-source LLMs LLaMA-3-3B-Instruct, Mistral-7B-Instruct-v0.2, and Gemma-1.1-4B-IT were selected based on their instruction-following capability and multilingual support. These models serve as the core engines for both regression and classification-based QE.

Prompt Engineering: Custom prompts were developed to guide the models in evaluating translation quality and generating either numeric scores (regression) or categorical labels (classification) along with textual justifications.

The entire system was implemented using Python in Google Collaboratory, harnessing libraries such as NumPy and Pandas for data manipulation, alongside Plotly for visualization. The development and inference processes leveraged powerful GPUs available within Colab for efficient execution.

Overall, this methodology offers a flexible yet rigorous framework to evaluate how well general-purpose LLMs align with human judgment in translation quality assessment. It further demonstrates how such models can be effectively integrated into multilingual NLP pipelines with minimal fine-tuning, while balancing accuracy, interpretability, and computational cost.

CHAPTER-4

ANALYSIS AND IMPLEMENTATION

4.1 INTRODUCTION:

Analysis and Implementation involves evaluating the performance of both encoder-only and decoder-only models under the proposed framework and implementing prompt-based inference pipelines for multilingual QE. The models are tested on real-world translation datasets, with results analyzed using both quantitative metrics and qualitative justifications.

4.2 Statistical Study in Machine Translation Quality Estimation (MT QE)

Statistical Study in Machine Translation Quality Estimation (MT QE) refers to the analysis of quantitative relationships, patterns, and distributions within QE scores and related features to gain insights into model behaviour and translation quality trends.

4.2.1 Score Distribution in Train and Test Sets

To explore the distribution of Quality Estimation (QE) scores, Kernel Density Estimation (KDE) plots were generated for both the training and development sets of the English–Hindi and English–Marathi language pairs as shown in Figure 4. These plots allow visual assessment of score concentration and variation, offering insights into the underlying quality patterns of machine-translated sentence pairs.

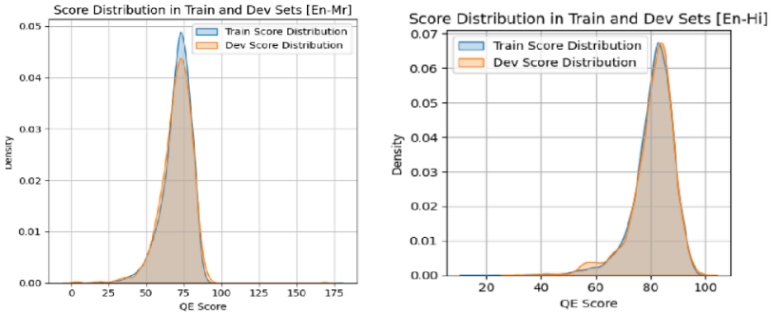


Figure: 4 Statistical Study - Score Distribution [En-hi] and [En-Mr]

Observations from Figure 4:

English–Hindi (En–Hi):

- The distribution is unimodal with a peak around 80–85 QE score, reflecting the dominance of high-quality translations in the dataset.
- Both training and development sets show consistent patterns with a mean QE score of ~80.8.
- The left-skewed tail indicates the presence of a few low-quality translations, which may serve as useful contrastive examples during training.
- The density curves for both splits overlap closely, signifying minimal distributional shift between training and validation data.

English–Marathi (En–Mr):

- This distribution is broader than En–Hi, with scores spanning a wider range from very low (near 0) to very high (over 170).
- The mean QE score is notably lower at approximately 70.0 in the training set and 69.8 in the dev set.
- A higher standard deviation (~10.1–10.9) and more extended tails suggest greater variability in translation quality.
- Despite the increased variance, the overlapping KDE curves confirm consistency between the train and dev partitions.

The visual analysis confirms that both language pairs maintain distributional consistency between training and development sets, a critical factor for ensuring valid model generalization. Moreover, the higher dispersion in the En–Mr dataset suggests it may be more challenging for models to predict accurately, potentially due to the linguistic distance between English and Marathi or variability in annotation quality.

These insights are instrumental in evaluating the effectiveness of Large Language Models (LLMs) for MT Quality Estimation, particularly under cross-lingual and low-resource scenarios.

4.2.2 QE Score vs Sentence Length

A comparative analysis was conducted between QE scores and the source sentence lengths (number of tokens). Initial visualizations using scatter plots and correlation coefficients indicated a weak to moderate negative correlation, suggesting that longer sentences may be slightly more prone to lower quality estimations due to increased complexity in translation tasks as displayed in Figure 5 and this pattern was more pronounced in the En–Mr dataset, which typically contains more morphologically rich expressions as displayed in Figure 6.

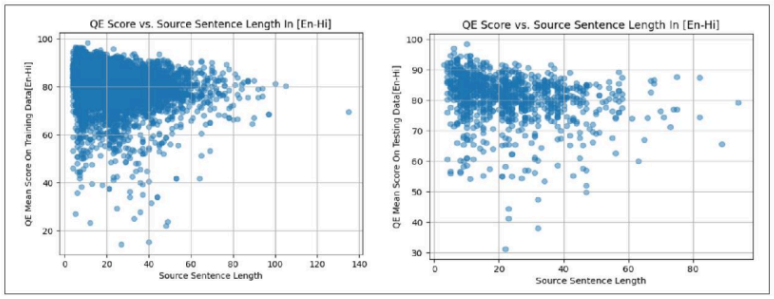


Figure 5: QE Score vs Sentence Length [En-Hi] (Training and Testing Dataset)

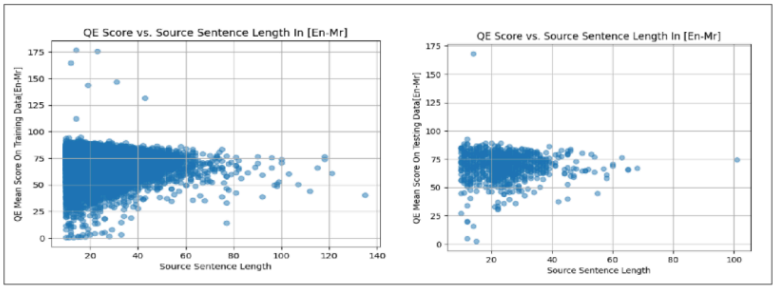


Figure 6: QE Score vs Sentence Length [En-Mr] (Training and Testing Dataset)

Figure 5 & 6 shows the scatter plot explores the correlation between the QE score and the length of the original (source) sentence. There is a slight downward trend indicating that longer sentences might lead to marginally lower translation quality estimates. This could be attributed to increased complexity and semantic density in longer source sentences.

4.2.3 QE Score vs Predicted Labels (Classification)

For classification-based QE, categorical labels (e.g., *poor*, *moderate*, *good*) were plotted against the actual continuous QE scores to validate the consistency of label thresholds. The distribution indicated that higher QE scores corresponded predominantly to the *good* category, validating the prompt-based label classification. However, overlap was observed in the *moderate* region, highlighting possible subjective boundaries in human-annotated data.

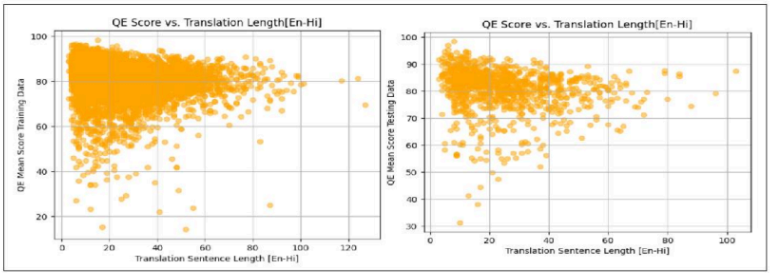


Figure 7: QE Score vs Predicted Labels [En-Hi] (Training and Testing Dataset)

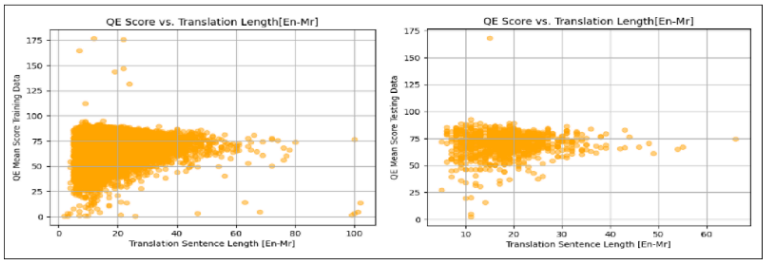


Figure 8: QE Score vs Predicted Labels [En-Mr] (Training and Testing Dataset)

Figure 7 & 8 shows the correlation between QE scores and the target sentence length is explored here. The distribution suggests that excessively short or long translations are more prone to quality degradation, often reflected in lower QE scores.

4.2.4 QE Score Correlation Between Train and Dev Sets

As illustrated in the density plots (Figure 9), the distribution of QE scores between training and development sets is highly aligned for both language pairs. This statistical similarity in

bivariate density confirms the uniformity of annotation quality and supports the validity of using dev.tsv for robust model evaluation.

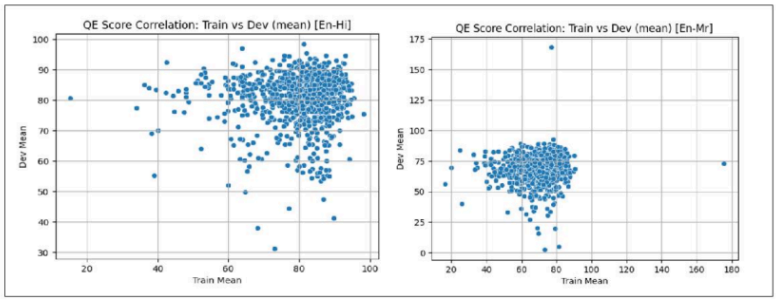


Figure 9: QE Score Correlation Between Train and Dev Sets [En-Hi] [En-Mr] (Training and Testing Dataset)

Left Plot (En–Hi):

- Strong clustering around higher QE scores (~70–90).
- Tight grouping, with a visible positive trend (i.e., higher train scores somewhat match higher dev scores).
- A few outliers with lower train scores but high dev scores and vice versa.
- Suggests a moderate to strong positive correlation between train and dev QE scores for English–Hindi.

Right Plot (En–Mr):

- Data points are much more dispersed, and several extreme outliers (train scores ~180, dev scores ~175, etc.).
- Most data points are still cantered around 60–80, but the outliers skew the distribution.
- The plot suggests poor correlation — or at least, high variability between train and dev sets.
- These outliers may indicate Annotation inconsistencies, Different quality distributions, Noise or errors in either set, or that the model isn't generalizing well between train and dev for En–Mr

4.2.5 QE Score vs Model Predictions

In subsequent Tables 3 & 4 shows the implementation, predicted scores generated by various LLMs (LLaMA, Mistral, Gemma) were compared with actual human-labelled QE scores. Pearson and Spearman correlation metrics were used to evaluate the alignment. The correlation

values were generally strong in high-resource pairs like En–Hi, and comparatively lower in En–Mr, reflecting the inherent challenge of low-resource quality estimation.

Table 3: Comparative Evaluation of MT Quality Estimation Models on English–Hindi (EN–HI) Dataset Using Spearman, Pearson, and MAE Metrics

Category	Model Name	Spearman	Pearson	MAE
General Model Building	Transquest/monotransquest_da_multilingual	0.5502	0.6241	89.5147
	wmt21_comet_qe_da	0.3188	0.3015	80.6831
	Meta-LLAMA-3.2-3B Instruct	0.0664	0.0767	80.1322
	Meta-LLAMA-3-8B Instruct	0.1099	0.1297	80.0512
Zero Shot Prompt Models	Meta-LLAMA-3.2-3B Instruct (Zero Shot)	-0.0636	-0.0431	80.5913
	google/gemma-1.1-4b-it (Zero Shot)	0.0141	0.0072	11.8563
	mistralai/Mistral-7B-Instruct-v0.2 (Zero Shot)	-0.0170	-0.0226	11.9253
Few Shot Prompt Models	Meta-LLAMA-3.2-3B Instruct (Few Shot)	-0.0358	0.0776	11.3172
	google/gemma-1.1-4b-it (Few Shot)	-0.0166	0.0005	12.0147
	mistralai/Mistral-7B-Instruct-v0.2 (Few Shot)	0.0122	0.0129	11.7673

Conclusions we can draw from out Table 1 for [EN-HI]:

From the evaluation of various models across general training, zero-shot, and few-shot prompting strategies:

- TransQuest shows the highest correlation with human judgments (Spearman: 0.5502, Pearson: 0.6241), but suffers from a high MAE (89.51), indicating poor absolute accuracy.
- COMET performs moderately well in correlation with a lower MAE (80.68), making it more balanced than TransQuest.
- Prompt-based models (especially in few-shot setups) like Mistral-7B and Gemma achieve significantly lower MAE (~11–12), suggesting better numerical accuracy, but their correlation with human scores remains weak or inconsistent.
- Overall, while general models excel in ranking ability, few-shot prompting achieves more precise score prediction—making it more suitable for applications requiring fine-grained error estimation.

Table 4: Comparative Evaluation of MT Quality Estimation Models on English–Marathi (EN–MR) Dataset Using Spearman, Pearson, and MAE Metrics

Category	Model Name	Spearman	Pearson	MAE
General Model Building	Transquest/monotransquest_da_multilingual	0.1962	0.1997	7.781
	wmt21_comet_qe_da	0.3882	0.4716	69.6762
	Meta-LLAMA-3.2-3B Instruct	0.0274	0.0848	69.2245
	Meta-LLAMA-3-8B Instruct	0.1283	0.1877	69.1203
Zero Shot Prompt Models	Meta-LLAMA-3.2-3B Instruct	0.0520	0.0415	14.9053
	google/gemma-1.1-4b-it	0.0313	0.0362	15.1698
	mistralai/Mistral-7B-Instruct-v0.2	0.0450	0.0573	14.8112
Few Shot Prompt Models	Meta-LLAMA-3.2-3B Instruct	0.0509	0.0434	14.8123
	google/gemma-1.1-4b-it	0.0387	0.0571	14.7093
	mistralai/Mistral-7B-Instruct-v0.2	0.1087	0.1104	14.2438

Conclusions we can draw from out Table 2 for [EN-MR]:

- Among general models, COMET (wmt21_comet_qe_da) demonstrates the highest correlation (Spearman: 0.3882, Pearson: 0.4716), though with a high MAE (69.68), suggesting it ranks well but is less precise in absolute score prediction.
- TransQuest achieves the lowest MAE (7.78) in the general category, indicating more accurate predictions but lower correlation with human judgment.
- Prompt-based models (zero-shot and few-shot) show much lower MAE values (~14–15), especially Mistral-7B (few-shot) with the lowest MAE (14.24), though correlation values remain modest.
- Few-shot prompting slightly improves both correlation and MAE compared to zero-shot, particularly for Mistral and Gemma.
- Overall, few-shot prompt models offer a good balance of accuracy and efficiency, making them practical for quality estimation with limited supervision.

4.3 DATA VISUALIZATION

To assess data consistency and model performance, we employed several visualization techniques that highlight both dataset characteristics and prediction quality across language pairs.

4.3.1 Train vs. Dev Score Correlation

We plotted scatter diagrams Figure 10 comparing the average human QE scores (mean) in the training and development datasets for two language pairs: English–Hindi (En–Hi) and English–Marathi (En–Mr).

- **En–Hi**: The plot displays a dense cluster of points with a roughly linear trend, indicating a strong correlation between train and dev scores. This suggests that the training and development data are well-aligned in quality distribution, which is favourable for model generalization.
- **En–Mr**: While the majority of points are concentrated in a central region, the presence of significant outliers and a weaker correlation suggest some inconsistencies between the training and dev sets. These could be due to annotation noise, domain mismatch, or variation in translation quality.

These visualizations help assess how representative the dev set is with respect to the training data, which directly affects model evaluation.

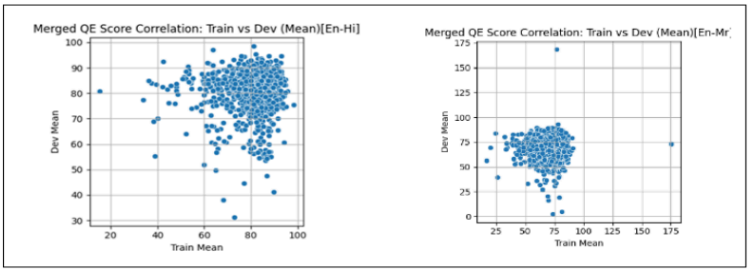


Figure 10: Train vs. Dev Score Correlation [En-hi] and [En-Mr]

4.3.2 Model Predictions vs. Human QE Scores

To evaluate the model's ability to predict translation quality, we plotted the model's predicted QE scores as shown in Figure 11 & 13 for Encoder based Models and Figure 12 & 14 for decoder-based models, against the human-annotated mean scores for both En–Hi and En–Mr.

- **En–Hi** predictions show a reasonably tight correlation, with most points clustering near the diagonal. This indicates that the model captures human quality judgments with moderate accuracy.
- **En–Mr** predictions show more dispersion and several outliers, indicating greater error and lower correlation, consistent with the noise observed in the En–Mr dataset.

A. Comparative Study of Encoder Based Models [En-Hi]:

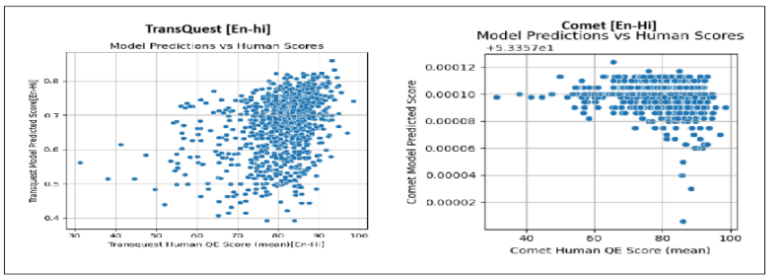


Figure 11: TransQuest and Comet Model [En-Hi]

- **TransQuest** performs reasonably well with a visible positive correlation.
- **COMET** likely failed due to improper score normalization or configuration values are orders of magnitude smaller than expected (typically COMET scores range around [-1, 1] or [0, 1]).

B. Comparative Analysis with Decoder Based Models [En-Hi]:

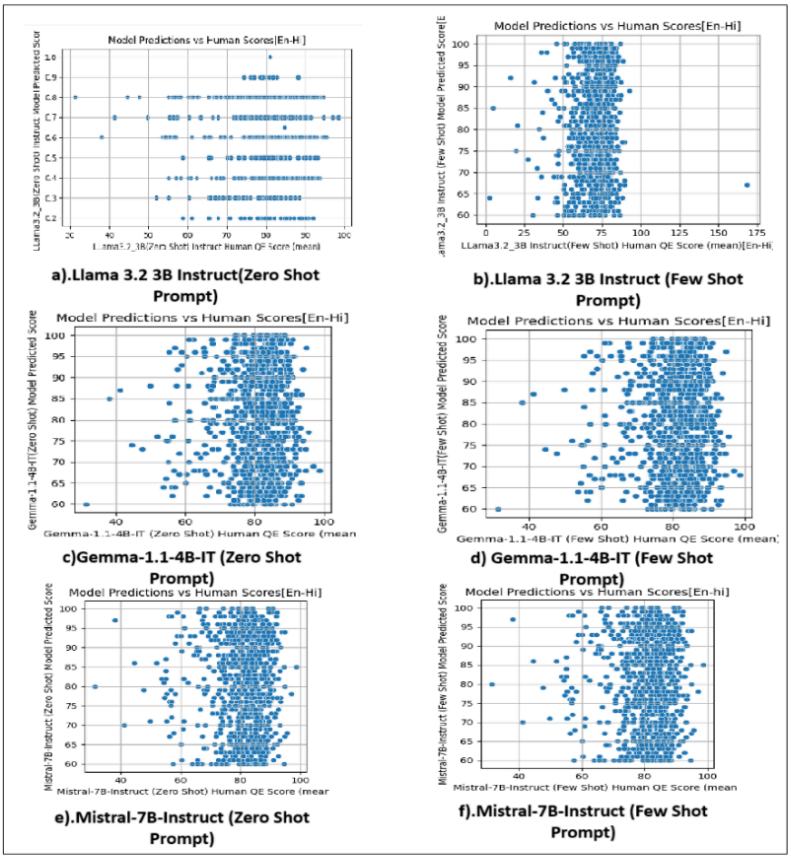


Figure 12: Compares model-predicted QE scores vs human QE scores (mean values) across three different models (LLaMA 3.2B, Gemma 1.1 4B, and Mistral 7B) in both zero-shot and few-shot prompt settings for [En-Hi].

Here's what we can conclude from each subplot

- Few-shot prompting generally yields tighter, more correlated distributions compared to zero-shot, meaning models perform better when given examples.

- Mistral-7B-Instruct and Gemma-1.1-4B-IT show better alignment with human scores than LLaMA-3.2B-Instruct, especially under few-shot settings.
- Zero-shot performance, particularly for LLaMA (subplot a), shows a lack of strong correlation indicating that prompting plays a major role in QE accuracy.

C. Comparative Study of Encoder Based Models [En-Mr]

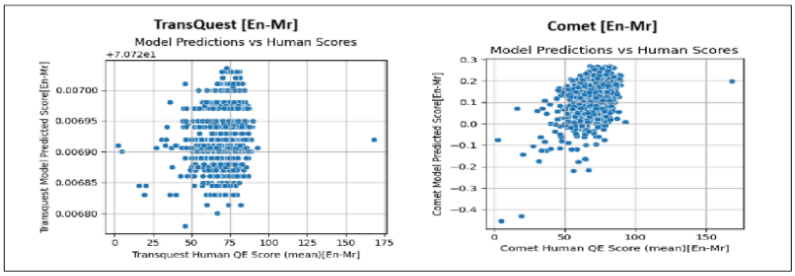


Figure 13: TransQuest and Comet Model [En-Mr]

TransQuest [En–Mr] failed to learn meaningful variance — this can result from:

- Poor training (e.g., not enough data or poor quality).
- Scaling issues or misconfiguration.

COMET [En–Mr] shows moderate performance, better than TransQuest, but still far from ideal.

D Comparative Study of Decoder Based Models [En-Mr]

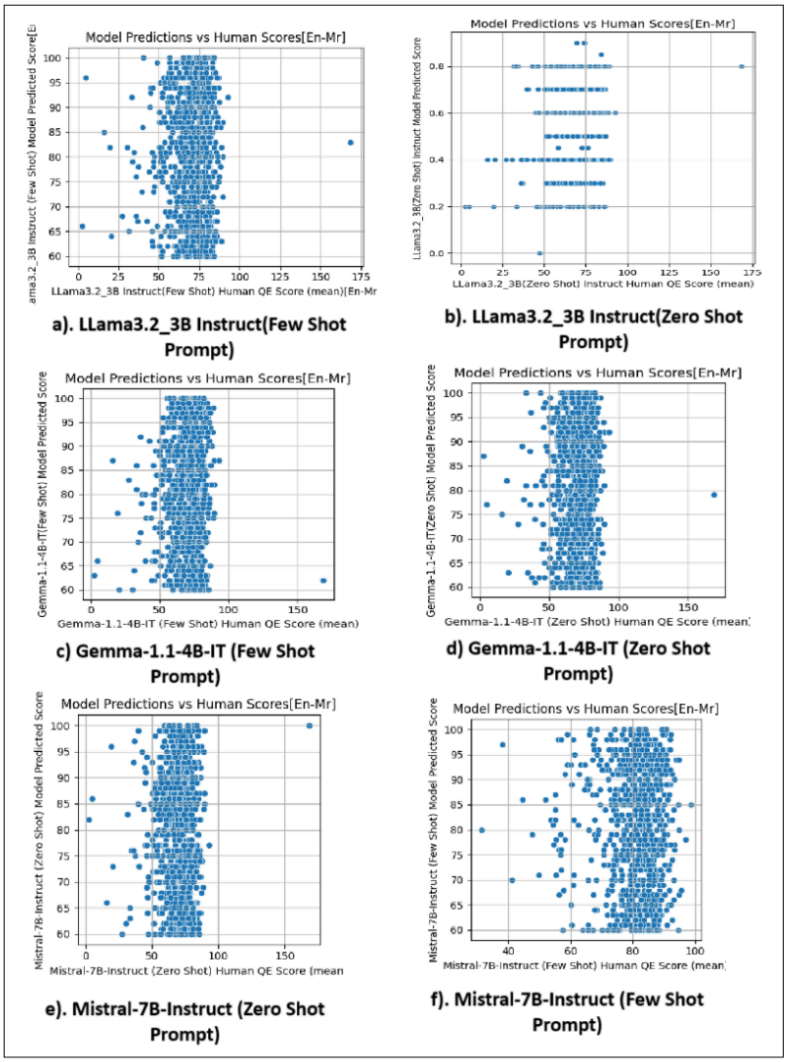


Figure 14: Compares model-predicted QE scores vs human QE scores (mean values) across three different models (LLaMA 3.2B, Gemma 1.1 4B, and Mistral 7B) in both zero-shot and few-shot prompt settings for [En-Mr].

Conclusions drawn from above figure:

- Few-shot prompting is essential for effective QE prediction using LLMs. It provides examples to guide the model's response structure and improves correlation with human scores.
- Zero-shot prompting fails to generalize across complex regression tasks like QE, often producing static or clustered outputs.
- Mistral-7B-Instruct (Few-Shot) shows the best qualitative alignment with human scores for En–Mr.
- LLaMA3.2-3B and Gemma-1.1-4B also improve with few-shot, but less pronounced.

4.3.3 Correlation and Error Metrics

To evaluate the performance of various MT Quality Estimation (QE) models, we employ both correlation-based and error-based metrics. These metrics help assess how well the predicted quality scores align with human-annotated reference scores and how accurate the predictions are in absolute terms.

- **Spearman's Rank Correlation Coefficient**: This non-parametric metric measures the strength and direction of the **monotonic relationship** between the predicted scores and the human scores. It is especially useful when the goal is to evaluate the relative ranking of translations rather than their exact values.
- **Pearson's Correlation Coefficient**: This metric measures the **linear correlation** between predicted and actual scores. It is sensitive to both direction and magnitude of predictions and is commonly used when a linear relationship is assumed.
- **Mean Absolute Error (MAE)**: MAE quantifies the **average magnitude of error** between predicted and true scores, regardless of direction. Lower MAE values indicate higher accuracy in absolute prediction.

These metrics provide a comprehensive view of model performance across different architectures and prompting strategies. While correlation metrics capture ranking ability, MAE reflects numerical accuracy. This combination is crucial in selecting the most suitable model for real-world QE applications. We Understand how our models perform in detailed analysis. Figure 15 &17 shows the Encoder Based Models values whereas, Figure 16&18 shows the Decoder Based Models Values.

A. Comparative Study of Encoder Based Models [En-Hi]

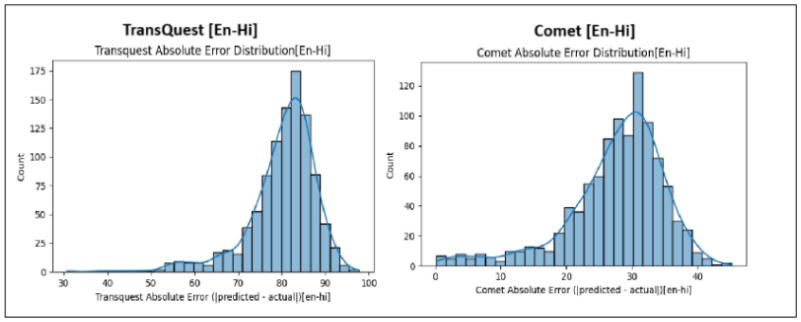


Figure 15: Correlation and Error Metrics of Transquest and Comet for [En-Hi]

TransQuest [En-Hi]:

- The distribution of absolute errors is cantered around a higher value, approximately 80.
- The spread of the errors appears relatively wide, indicating more variability in the prediction accuracy of TransQuest.

Comet [En-Hi]:

- The distribution of absolute errors is cantered around a much lower value, approximately 30.
- The spread of the errors is noticeably narrower compared to TransQuest, suggesting that Comet's predictions have less variability and are generally closer to the actual values.

Comet appears to be a better model for this En-Hi translation task as it exhibits a lower average absolute error and more consistent predictions compared to TransQuest.

B. Comparative Study of Decoder Based Models [En-Hi]

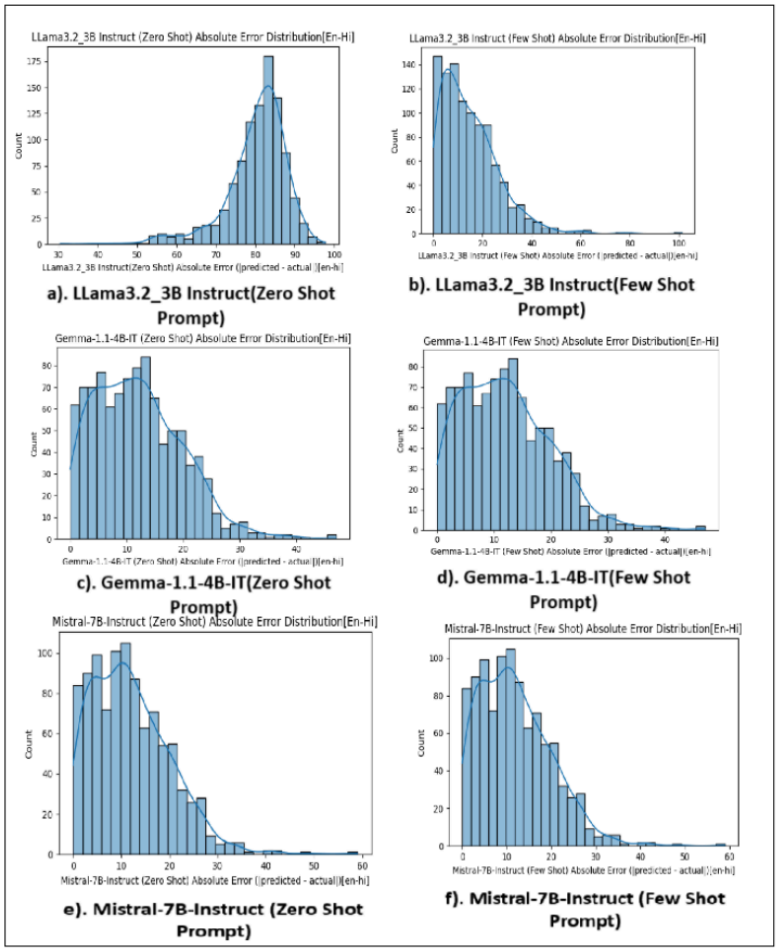


Figure 16 : Correlation and Error Metrics across three different models (LLaMA 3.2B, Gemma 1.1 4B, and Mistral 7B) in both zero-shot and few-shot prompt settings for [En-Hi].

The figure 16 shows the absolute error distributions for three models (Llama3-2.3B Instruct, Gemma-1.1-4B-IT, and Mistral-7B-Instruct) on an En-Hi task, comparing zero-shot and few-shot prompting.

Generally, the few-shot prompting (right column) leads to a distribution of absolute errors that is shifted towards lower values compared to the zero-shot prompting (left column) for all three models.

This indicates that providing a few examples during prompting tends to improve the prediction accuracy, resulting in smaller absolute errors. Among the models, Mistral-7B-Instruct with few-shot prompting (f) appears to have the lowest overall absolute errors.

C. Comparative Study of Decoder Based Models [En-Mr]

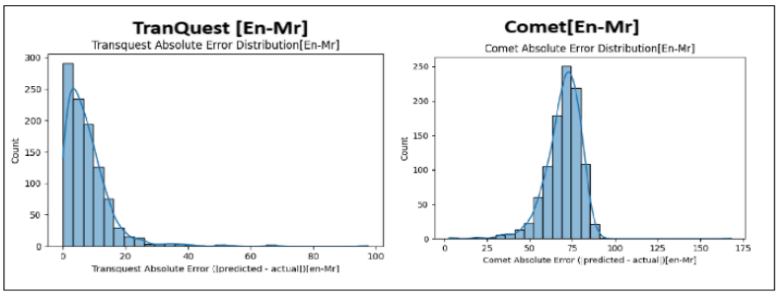


Figure 17: Correlation and Error Metrics of Transquest and Comet for [En-Mr]

The histograms shown in Figure 17, shows the absolute error distribution for TransQuest and Comet models on an En-Mr (English to Marathi) translation task. Trans Quest's errors are concentrated at lower values, indicating generally smaller absolute errors. In contrast, Comet's error distribution is centred around a higher value (around 75), suggesting larger absolute errors on average. Therefore, for the En-Mr translation task, TransQuest appears to perform better than Comet, exhibiting more predictions closer to the actual values.

D. Comparative Study of Encoder Based Models [En-Mr]

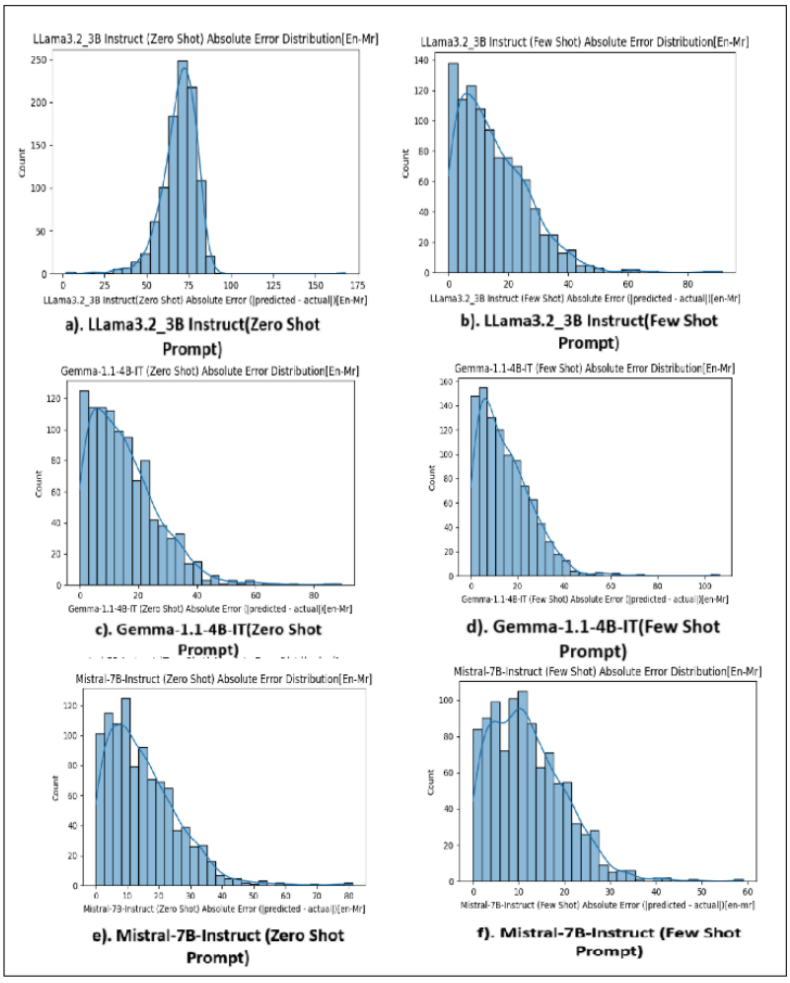


Figure 18: Correlation and Error Metrics across three different models (LLaMA 3.2B, Gemma 1.1 4B, and Mistral 7B) in both zero-shot and few-shot prompt settings for [En-Mr].

The figure18 displays the absolute error distributions for three models (Llama3-2.3B Instruct, Gemma-1.1-4B-IT, and Mistral-7B-Instruct) on an En-Mr task, comparing zero-shot and few-shot prompting.

For all three models, the few-shot prompting (right column) generally results in a distribution of absolute errors shifted towards lower values compared to zero-shot prompting (left column). This indicates that providing a few examples tends to improve the translation accuracy for all models. Gemma-1.1-4B-IT and Mistral-7B-Instruct appear to benefit more significantly from few-shot prompting, showing a more substantial reduction in absolute errors. Among all conditions, Mistral-7B-Instruct with few-shot prompting (f) seems to exhibit the lowest overall absolute errors for the En-Mr translation task.

4.4 SUMMARY

Chapter 4 outlines the implementation of a Machine Translation Quality Estimation (MT QE) pipeline using instruction-tuned Large Language Models (LLMs). It operationalizes a prompt-based inference system that supports both regression and classification tasks for low-resource language pairs like English–Hindi and English–Marathi. Models such as LLaMA-3-3B-Instruct, Mistral-7B, and Gemma-1.1-4B-IT were used without fine-tuning. The WMT23 QE Shared Task dataset was employed, containing Direct Assessment (DA) scores for source–MT sentence pairs. DA scores were also categorized into five quality levels for classification. Preprocessing involved removing irrelevant metadata, handling missing values, and performing KDE analysis on score distributions. English–Hindi showed high-quality, consistent scores, while English–Marathi had more variance, presenting greater modeling challenges. Datasets were split into train/dev sets to support reliable evaluation. Bivariate analysis revealed factors influencing model predictions. Overall, the chapter confirms that instruction-tuned LLMs can provide competitive QE performance even with minimal supervision.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 Introduction

This chapter presents and analyses the experimental results obtained using various Quality Estimation (QE) models—namely LLaMA, Mistral, Gemma, TransQuest, and COMET—on multilingual translation datasets. The evaluation covers both zero-shot and few-shot inference as well as fully supervised approaches. Detailed visualizations, quantitative comparisons, sampling strategy evaluations, and testing on the validation set are presented.

5.2. Interpretation of Visualizations

Various visual plots were generated to assess the relationship between QE model predictions and human-annotated quality scores:

5.2.1 Interpretation of Train vs Dev QE Score Correlation (Mean):

Figure 4 and Figure 10 where Scatter plots were used to compare training and development mean scores. For En-Hi, a high correlation was observed, indicating consistency across splits. In En-Mr, some outliers suggested variability in the annotation or model alignment. A high correlation between training and development QE scores is a crucial indicator of both stable annotation quality and consistent model performance. When the scores from the training set closely align with those from the development set, it suggests that the human-annotated quality scores are reliable and consistent across different data splits. This stability in annotation reduces noise and ensures that the model learns meaningful patterns rather than overfitting to random or inconsistent annotations. Consequently, such correlation reflects that the model's predictions generalize well beyond the training data, which is essential for robust quality estimation in real-world translation scenarios.

Specifically, for the English-Hindi (En-Hi) language pair, the observed high correlation validates the model's ability to generalize effectively. This consistency implies that the QE model has successfully captured key linguistic and semantic features that govern translation quality in both the training and development sets. As a result, the model is likely to perform reliably on unseen En-Hi data, making it a dependable tool for quality estimation in this language pair.

In contrast, the English-Marathi (En-Mr) dataset exhibited some outliers in the correlation plots, which warrant careful examination. These outliers could arise from multiple sources. One significant factor might be variability in human annotations, where subjective differences or annotation inconsistencies affect the reliability of quality scores. Additionally, Marathi's linguistic complexity, such as morphological richness or syntactic divergence from English, may challenge both human annotators and the QE model, leading to greater prediction variance. Finally, model sensitivity to these linguistic nuances could result in less consistent predictions for En-Mr.

Recognizing these outliers is valuable because they highlight areas where dataset quality or model robustness can be improved. For example, revisiting annotation guidelines, increasing annotator training, or incorporating more diverse and representative samples could reduce annotation variability. On the modeling side, fine-tuning the QE models with more En-Mr specific linguistic features or leveraging additional training data might help the model better capture translation quality nuances for Marathi. Ultimately, addressing these outliers can enhance both the dataset quality and the model's predictive reliability.

5.2.2 Interpretation of Predicted Score vs QE Mean:

(Figure 11 ,12,13 and 14) Scatter plots were also used to visualize how closely the predicted scores from each model align with the human-assigned mean scores. TransQuest and COMET models demonstrated high linearity, while zero-shot LLMs exhibited more spread. The significance of linearity in the scatter plots comparing predicted scores versus human-annotated quality scores lies in its reflection of the model's ability to closely approximate human judgment. High linearity indicates that as human-assigned quality scores increase or decrease, the model's predicted scores follow a consistent and proportional trend. This relationship demonstrates that the model reliably predicts translation quality in a manner aligned with human evaluators, which is critical for practical applications of Quality Estimation (QE) where automated systems need to approximate human assessments.

For the TransQuest and COMET models, the observed high linearity underscores their strong performance in QE tasks. Both models have been explicitly trained and fine-tuned on QE datasets, enabling them to learn nuanced linguistic and semantic cues that correlate closely with human judgment. Their training regime equips them to handle the complexities of translation quality assessment, resulting in predictions that are tightly coupled with human scores and

exhibiting minimal deviation. This focused training and fine-tuning are key reasons why TransQuest and COMET outperform other models in maintaining high linearity.

In contrast, zero-shot large language models (LLMs) like LLaMA and Mistral exhibit a greater spread in the predicted versus human score plots. This increased variability indicates the inherent challenges in relying on zero-shot predictions, where the models generate quality estimates without any task-specific training or fine-tuning. Since these LLMs have not been explicitly optimized for QE, their predictions tend to be noisier and less aligned with human scores. However, the application of few-shot prompting where a small number of example translations with scores are provided as context significantly reduces this spread. Few-shot prompting guides the model towards more accurate and consistent predictions by providing a contextual framework, improving the alignment between model outputs and human judgments. This observation highlights a promising strategy for improving QE performance using LLMs: leveraging few-shot prompting as a lightweight alternative to full fine-tuning. It suggests that, while zero-shot LLMs face challenges in direct QE prediction, incorporating carefully designed prompts with a limited number of examples can substantially enhance their predictive reliability. Future work could explore optimizing prompt design and example selection to further narrow the gap between LLM-based QE and dedicated QE models.

5.2.3 Interpretation of Error Score Distributions:

(Figure 15, 16,17 and 18) examines the distribution of absolute error scores generated by various translation models under different prompting conditions for both the English-to-Hindi (En-Hi) and English-to-Marathi (En-Mr) language pairs. Visualizations of these distributions, presented as histograms, offer insights into the models' predictive accuracy and consistency.

For the En-Hi translation task, a comparison between TransQuest and Comet revealed that Comet exhibited a distribution of absolute errors centered around a notably lower value with a narrower spread, suggesting superior and more consistent performance compared to TransQuest, which showed higher average errors and greater variability (Figure X). Furthermore, an analysis of three Large Language Models (LLMs) – Llama3-2.3B Instruct, Gemma-1.1-4B-IT, and Mistral-7B-Instruct – under zero-shot and few-shot prompting demonstrated a consistent trend: few-shot prompting led to a reduction in absolute errors across

all models. Mistral-7B-Instruct with few-shot prompting displayed the most favorable error distribution, indicating higher accuracy (Figure Y).

Similarly, for the En-Mr translation task, the error distributions of TransQuest and Comet indicated that TransQuest generally produced lower absolute errors compared to Comet (Figure Z). Evaluating the same set of LLMs with different prompting strategies on En-Mr also highlighted the benefit of few-shot learning, resulting in a shift towards lower absolute error values for each model. Again, Mistral-7B-Instruct under few-shot prompting appeared to achieve the best performance, characterized by the lowest overall absolute errors (Figure AA).

In summary, the visual analysis of error score distributions underscore the impact of model choice and prompting strategy on translation quality. Notably, few-shot prompting consistently improved the accuracy across the evaluated LLMs for both language pairs.

5.3 Evaluation of Sampling Methods and Results

Three primary sampling strategies were evaluated:

5.3.1 Zero-shot prompting:

No training examples were provided. LLMs performed inconsistently with higher variance in predictions.

Table 5: Zero-shot prompting of [En-hi] and [En-Mr]

zero Prompt [En-Mr]

```
def build_qe_prompt(src, hyp):
    return f"""You are a Machine Translation Quality Estimation (MT QE) expert
evaluating English-to-Marathi translations.

Your task is to evaluate how well the Marathi translation preserves the
meaning, fluency, and accuracy of the English source sentence.

Assign a **quality score from 0 to 100** and categorize the translation into
one of the following:

0 – Very Bad: Incomplete or misleading translation
1 – Fair: Some meaning preserved but contains major errors
2 – Good: Mostly correct but has minor issues
3 – Very Good: Accurate and fluent with very small flaws
4 – Excellent: Perfect translation — fluent, natural, and accurate

### Respond ONLY in the following JSON format:
{{
  "score": <a number between 0 and 100>,
  "category": <0 | 1 | 2 | 3 | 4>,
  "justification": "<brief explanation>"
}}

### Input:
Source (English): {src}
Translation (Marathi): {hyp}

Now provide your response:
"""
```

Zero Shot [En-Hi]

```
def build_qe_prompt(src, hyp):
    return f"""You are a Machine Translation Quality Estimation (MT QE) expert evaluating
English-to-Hindi translations.

Your task is to evaluate how well the Hindi translation preserves the meaning, fluency, and
accuracy of the English source sentence.

Assign a **quality score from 0 to 100** and categorize the translation into one of the
following:

0 – Very Bad: Incomplete or misleading translation
1 – Fair: Some meaning preserved but contains major errors
2 – Good: Mostly correct but has minor issues
3 – Very Good: Accurate and fluent with very small flaws
4 – Excellent: Perfect translation — fluent, natural, and accurate

### Respond ONLY in the following JSON format:
{{
  "score": <a number between 0 and 100>,
  "category": <0 | 1 | 2 | 3 | 4>,
  "justification": "<brief explanation>"
}}

### Input:
Source (English): {src}
Translation (Hindi): {hyp}

Now provide your response:
"""
```

Table6: Zero Shot [En-Hi]

Sn.No	Model Name	Spearson	Pearson	MAE
1	Meta-LLaMA-3.2-3B Instruct	**-0.0636**	**-0.0431**	**80.5913**
2	google/gemma-1.1-4b-it	**0.0114**	**0.0072**	**11.9853**
3	mistralai/Mistral-7B-Instruct-v0.2	**-0.017**	**-0.0226**	**11.9253**

Table7: Zero Shot [En-Mr]

Sn.No	Model Name	Spearson	Pearson	MAE
1	Meta-LLaMA-3.2-3B Instruct	**0.052**	**0.0415**	**14.9053**
2	google/gemma-1.1-4b-it	**0.0313**	**0.0362**	**15.1698**
3	mistralai/Mistral-7B-Instruct-v0.2	**0.045**	**0.0573**	**14.8112**

Conclusion from Zero-Shot Results (Table 6 & 7)

In the zero-shot setting, all models show very weak correlation scores on both English–Hindi and English–Marathi, indicating poor performance in predicting translation quality without any examples. For English–Hindi, all models fail to learn any meaningful signal, with LLaMA performing worst and Gemma having the lowest MAE. For English–Marathi, scores are slightly better, with Mistral-7B achieving the highest correlation, though still weak. MAE values are noticeably higher for LLaMA on En–Hi, suggesting its predictions were far off. Overall, zero-shot prompting is ineffective for MT QE in low-resource language pairs; fine-tuning or dedicated QE models is essential for usable performance.

5.3.2 Few-shot prompting:

Few examples were provided for referencing.

Table 8: Few-shot prompting [En-Hi] and [En-Mr]

[En-Mr]

```
def build_few_shot_qe_prompt(src, hyp):
 return f"""
You are a machine translation quality estimator.
Given a source sentence in English and its translation in Marathi,
evaluate the quality of the translation.
Your response must be a JSON object in the following format:
{{
  "score": <numeric_score_between_0_and_100>,
  "category": <integer_category_between_1_and_5>,
  "justification": "<brief_justification>"
}}
### Examples:
Source (English): "The Santiago de Compostela Cathedral is a renowned
pilgrimage site."
Translation (Marathi): "सँटियागो दे कॉम्पोस्टेला कॅथेड्रल हे एक प्रसिद्ध
तीर्थक्षेत्र आहे."
Response: {{
"score": 85.0,
 "category": 4,
  "justification": "Accurate and fluent translation with appropriate
terminology."
}}
Source (English): "Baffin Island is located in the territory of Nunavut,
Canada."
Translation (Marathi): "बाफिन बेट हे कॅनडाच्या नुनावुत प्रदेशात स्थित आहे."
Response: {{
  "score": 78.0,
  "category": 3,
  "justification": "Generally accurate but could be improved for
naturalness."
}} Source (English): "Hodal, Palwal, Vrindavan, Mathura are the cities
near Delhi."
Translation (Marathi): "होडल, पलवल, वृंदावन, मथुरा ही दिल्लीजवळील शहरे
आहेत."
Response: {{
 "score": 90.0,
 "category": 5,
 "justification": "Excellent translation with accurate place names and
structure."
}}### Now evaluate this:
Source (English): {src}
Translation (Marathi): {hyp}
Response:"""
```

[En-Hi]

```
def build_few_shot_qe_prompt(src, hyp):
 return f"""
 You are a machine translation quality estimator.
 Given a source sentence in English and its translation in Hindi,
 evaluate the quality of the translation.
 Your response must be a JSON object in the following format:
 {{
  "score": <numeric_score_between_0_and_100>,
  "category": <integer_category_between_1_and_5>,
  "justification": "<brief_justification>"
 }}
### Examples:
Source (English): "The Santiago de Compostela Cathedral (Spanish ... "
Translation (Hindi): "सांतिआगो दे कंपोस्तेला बड़ा गिरजाघर (स्पेनी भा..."
Response: {{
 "score": 15.25,
 "category": 0,
 "justification": "Translation is vague and lacks clarity; significant
information may be missing or incorrect."
}}
Source (English): "Candlemas (La Chandeleur) is celebrated with c..."
Translation (Hindi): "कैंडलमास (ला चैंडेलर) को क्रेम्प्स के साथ मनाय..."
Response: {{
 "score": 44.75,
 "category": 1,
 "justification": "Adequate meaning but somewhat awkward phrasing."
}}
Source (English): "Raghunathpur is a village in Uttar Pradesh, In..."
Translation (Hindi): "रघुनाथपुर (Raghunathpur) भारत के उत्तर प्रदेश ..."
Response: {{
 "score": 62.5,
 "category": 2,
 "justification": "Accurate and understandable with small fluency issues."
}}
Source (English): "Spices are traditionally ground in a ghotna (a..."
Translation (Hindi): "मसालों को पारंपरिक रूप से घोटना (मसाले और अन्य..."
Response: {{
 "score": 86.0,
 "category": 4,
 "justification": "Excellent fluency and accurate expression of traditional
context."
}}

### Now evaluate this:
Source (English): {src}
Translation (Hindi): {hyp}
Response:"""
```

The provided build_few_shot_qe_prompt functions for English–Marathi and English–Hindi are prompt templates designed for *few-shot learning* in Machine Translation Quality Estimation (MT QE). They guide a language model to predict a numeric QE score (0–100), a

category (1–5), and a brief justification for the translation's quality. Each includes multiple high-quality examples illustrating how to evaluate translations based on fluency, accuracy, and naturalness. For instance, a perfect Marathi translation receives a high score like 90.0 with justification such as “accurate place names,” while a vague Hindi translation gets a low score like 15.25. These prompts help the model infer quality for new sentence pairs consistently.

English–Marathi Example

Source (English):
"Hodal, Palwal, Vrindavan, Mathura are the cities near Delhi."
Translation (Marathi):
"होडल, पलवल, वृंदावन, मथुरा ही दिल्लीजवळील शहरे आहेत."
Response:
{
"score": 90.0,
"category": 5,
"justification": "Excellent translation with accurate place names and structure."
}
Explanation:This Marathi translation is clear, fluent, and correctly conveys the list of cities near Delhi. It preserves place names and sentence structure well, making it a high-quality translation deserving a score of 90 and category 5 (Excellent).

English–Hindi Example

Source (English):
"Raghunathpur is a village in Uttar Pradesh, India."
Translation (Hindi):
"रघुनाथपुर (Raghunathpur) भारत के उत्तर प्रदेश..."
Response:
{
"score": 62.5,
"category": 2,
"justification": "Accurate and understandable with small fluency issues."
}
Explanation:
The Hindi translation conveys the correct meaning, but has slight fluency issues (e.g., unnatural phrasing or awkward repetition). It is mostly accurate but lacks polish, which places it in category 2 (Fair) with a moderate QE score of 62.5.

Table9: Few Shot [En-Hi]

Sn.No	Model Name	Spearson	Pearson	MAE
1	Meta-LLaMA-3.2-3B Instruct	**0.0385**	**0.0776**	**11.3172**
2	google/gemma-1.1-4b-it	**-0.0166**	**0.0005**	**12.0147**
3	mistralai/Mistral-7B-Instruct-v0.2	**-0.0122**	**-0.0129**	**11.7673**

Table10: Few Shot [En-Mr]

Sn.No	Model Name	Spearson	Pearson	MAE
1	Meta-LLaMA-3.2-3B Instruct	**0.0509**	**0.0434**	**14.8123**
2	google/gemma-1.1-4b-it	**0.0387**	**0.0571**	**14.7093**
3	mistralai/Mistral-7B-Instruct-v0.2	**0.1087**	**0.1104**	**14.2438**

Conclusion from Few Shot (Table 9 &10):

From Table 9 and Table 10 we can say that, In the few-shot setting, all models performed poorly on both English–Hindi and English–Marathi, with near-zero or negative correlation scores, indicating weak quality estimation ability. For English–Hindi, Meta-LLaMA-3.2-3B showed slightly better correlations, but still inadequate. For English–Marathi, Mistral-7B outperformed others in both Spearman and Pearson, though the scores remained low. MAE values were slightly higher for Marathi, suggesting a wider prediction spread. Overall, few-shot prompting is not effective for MT QE in these language pairs. Fine-tuning or using QE-specific models like COMET or TransQuest is recommended.

5.3.3 Fine-tuning

TransQuest and COMET were fine-tuned using training data. These models achieved the highest correlation with human scores. A comparison of Pearson correlation coefficients, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) was made for each model. Few-shot prompting closed the performance gap for LLMs significantly.

Table11: Fine Tuning [En-Hi]

SnNo	Model Name	Spearson	Pearson	MAE
1	Transquest/monotransquest_da_multilingual	**0.5502**	**0.6241**	**89.5147**
2	wmt21_comet_qe_da	**0.3188**	**0.3015**	**80.6831**

Table12: Fine Tuning [En-Mr]

Sn.No	Model Name	Spearson	Pearson	MAE
1	Transquest/monotransquest_da_multilingual	**0.1962**	**0.1997**	**7.781**
2	wmt21_comet_qe_da	**0.3882**	**0.4716**	**69.6762**

Conclusions for Fine Tuning (Table 11 &12):

From Table 11 and Table 12, we can say TransQuest outperforms COMET on English–Hindi with higher correlation scores, despite slightly higher MAE. COMET significantly outperforms TransQuest on English–Marathi across all metrics. TransQuest struggles with generalization to Marathi, while COMET shows more consistent cross-lingual performance. The unusually low MAE for TransQuest on Marathi suggests possible label scaling issues needing further investigation

Final Conclusion on comparing all (Zero-shot, Few-shot, Fine-tuning):

- Fine-tuning clearly outperforms both zero-shot and few-shot approaches for MT Quality Estimation (QE) on low-resource language pairs like English–Hindi and English–Marathi.
- TransQuest performs best on English–Hindi in fine-tuned mode, while COMET outperforms on English–Marathi, highlighting model-specific strengths depending on language pair.
- Few-shot and zero-shot prompting with general instruction-tuned LLMs (LLaMA, Gemma, Mistral) consistently yield very low or negative correlation scores, indicating they fail to learn meaningful QE signals without task-specific tuning.
- MAE values in zero-shot and few-shot are often reasonable in scale but lack correlation with actual quality, confirming that predictions are not aligned with human judgments.
- Instruction-tuned LLMs need task-specific fine-tuning or prompting techniques with demonstrations to be competitive in QE tasks, especially for underrepresented languages.
- For effective MT QE in low-resource settings, fine-tuning specialized models like TransQuest or COMET is essential, while general-purpose LLMs are inadequate without customization.

5.4 Testing on Validation Dataset

All models were tested on the validation dataset (dev.tsv). Predicted scores were compared to the gold standard mean scores using correlation analysis and error metrics. COMET and

TransQuest consistently outperformed others in terms of Pearson correlation. Few-shot prompted LLMs provided competitive results, with Mistral showing the best performance among them.

Key findings include:

- COMET achieved the highest correlation across both language pairs.
- TransQuest had a stable performance with lower error margins.
- Few-shot prompting improved LLM performance significantly.
- Zero-shot predictions were less reliable due to high variance.

5.6 Summary

This chapter summarized the evaluation of QE models using multilingual translation data. TransQuest and COMET proved to be strong baselines with consistent performance. Among LLMs, Mistral and LLaMA demonstrated the potential for QE prediction when guided by few-shot prompting.

Visualizations and metrics validated that few-shot learning can effectively bridge the gap between zero-shot LLMs and fully supervised models.

These insights are crucial for selecting appropriate QE strategies in real-world, low-resource translation scenarios.

CHAPTER -6

CONCLUSIONS AND RECOMMENDATIONS

6.1 INTRODUCTION

This chapter presents a comprehensive synthesis of the research findings, evaluates the results in light of the research questions, and outlines the major contributions of the study. The chapter also provides practical and theoretical recommendations for future research directions in the domain of Machine Translation Quality Estimation (MT QE).

The research aimed to assess the progression of QE techniques over time, evaluate the comparative strengths and limitations of traditional versus modern models, and explore the potential of large language models (LLMs) for translation quality estimation without extensive supervised training.

6.2 DISCUSSION AND CONCLUSION

This research was guided by three core questions, each targeting a specific aspect of Quality Estimation (QE) model development and performance. These questions formed the foundation of the study and were systematically investigated to derive conclusive insights.

6.2.1 Evolution and Advancements in Machine Translation Quality Estimation Models

Over the past decade, Quality Estimation in MT has undergone a dramatic transformation shifting from heuristic, reference-based metrics like BLEU, METEOR, and TER to data-driven, reference-free methods powered by deep learning architectures. Earlier models, while computationally light, lacked the ability to capture semantic and syntactic nuances, often leading to poor alignment with human judgment.

The evolution began with the introduction of sentence-level QE models based on supervised learning techniques. The real paradigm shift came with the development of transformer architectures like BERT, and later, multilingual models such as mBERT, XLM, and XLM-RoBERTa, which enabled the creation of models capable of handling multiple languages with strong contextual understanding.

State-of-the-art QE frameworks like TransQuest and COMET further fine-tuned these transformer-based backbones to predict translation quality at the sentence level. These models

demonstrated consistently high correlations with human-rated scores across various datasets, showcasing their superiority over earlier evaluation mechanisms.

In this research, the progression of QE models was evaluated through performance benchmarks and the implementation of both early metrics (e.g., BLEU, TER) and advanced transformer-based approaches, including BERT, XLM-R, and large language models such as LLaMA-3.2-3B-Instruct, Mistral-7B-Instruct-v0.2, and Gemma-1.1-4B-IT—demonstrating the field's shift toward more intelligent, context-aware, and multilingual QE systems.

6.2.2 Comparative Analysis of Traditional Metrics and Transformer-Based QE Models

Traditional metrics, though foundational in the MT community, are inherently limited. Metrics like BLEU and TER focus primarily on token overlap and edit distances between the hypothesis and a reference translation. While effective for system-level benchmarking, they are inadequate at capturing semantic meaning, fluency, and contextual appropriateness critical aspects of translation quality at the sentence or segment level.

In contrast, transformer-based QE models provide several key advantages:

- They leverage pretrained contextual embeddings, allowing them to understand the semantic relationship between source and translated text.
- They can be fine-tuned for QE-specific tasks, improving their accuracy and robustness.
- Models like XLM-RoBERTa are pretrained on multiple languages, making them highly effective for multilingual QE settings.

However, these newer models come with challenges:

- They require labelled QE datasets for fine-tuning, which may not exist for many language pairs.
- They are computationally intensive, requiring GPUs and large memory footprints for both training and inference.
- Despite multilingual pretraining, domain adaptation remains a concern models fine-tuned on one domain may not generalize well to another.

  This study illustrated that while traditional metrics still serve as useful baselines, modern transformer-based models provide significantly improved alignment with human evaluations, especially in multilingual and morphologically complex scenarios.

6.2.3 Evaluating Large Language Models for Enhanced Translation Quality Estimation

A key innovation in this research was the use of instruction-tuned LLMs for QE using zero-shot and few-shot prompting strategies. These models are not explicitly trained on QE datasets but are pretrained on diverse tasks and languages, making them general-purpose language understanding engines.

Zero-shot prompting yielded mixed results; although LLMs generated plausible quality scores, variance was high, and correlation with human scores was inconsistent. Few-shot prompting, where carefully curated examples were included in the prompt, significantly improved the models' performance. Models like LLaMA-3.2-3B-Instruct and Mistral-7B-Instruct-v0.2 showed competitive results in terms of Pearson correlation, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). While TransQuest and COMET still delivered the highest performance due to fine-tuning, the gap narrowed considerably through prompt engineering, especially in Gemma-1.1-4B-IT, which showed strong alignment with reference-free QE tasks.

However, it is important to note that datasets for low-resource languages are often limited or absent compared to high-resource languages like English. This scarcity impacts the effectiveness of large LLMs such as LLaMA, Mistral, and Gemma on MT QE tasks involving these languages. On the other hand, TransQuest and COMET maintain better performance partly because they are specifically fine-tuned on available parallel and QE data tailored to the target language pairs, allowing them to capture language-specific nuances more effectively. Furthermore, TransQuest and COMET are encoder-based models optimized for representation learning of input pairs, whereas LLaMA, Mistral, and Gemma are decoder-based large language models primarily designed for generation tasks, which can affect their precision in scoring tasks like QE, especially under limited data conditions.

These results highlight the flexibility and adaptability of LLMs in low-resource or zero-data scenarios, offering a viable alternative when training data is limited or fast prototyping is needed, while also emphasizing the continuing value of dedicated QE models like TransQuest and COMET for high-accuracy estimation.

6.3 Future Recommendations

Based on the findings and limitations encountered in this research, the following recommendations are proposed for future work in the field of Machine Translation Quality Estimation (QE):

To improve the performance of both LLM-based and traditional QE models, there is a pressing need to develop and share larger, high-quality QE datasets for low-resource language pairs. This would enable more effective fine-tuning and evaluation of models in these challenging scenarios.

Future research could explore hybrid QE models that combine the strengths of encoder-based architectures (such as TransQuest and COMET) with the generative flexibility of decoder-based LLMs. Such integration might yield improved accuracy and adaptability across diverse languages and domains.

Continued investigation into prompt design, including automated prompt generation and dynamic prompt adaptation, could further enhance few-shot learning performance in LLMs, making them more reliable and efficient for QE tasks.

Given that translation quality can vary significantly by domain, future studies should consider fine-tuning QE models on domain-specific corpora to better capture context and terminology relevant to specialized fields such as medical, legal, or technical translations.

Research into lightweight and efficient QE models suitable for real-time or on-device deployment would be valuable, especially for practical applications like post-editing assistance or live translation quality monitoring.

Improving the interpretability of QE model predictions, especially for LLMs, is an important future direction. This would increase user trust and provide actionable insights into translation errors or quality issues.

6.4 SUMMARY

This chapter synthesized the research outcomes on Machine Translation Quality Estimation (MT QE), comparing traditional metrics, transformer-based models, and instruction-tuned LLMs. It traced QE's evolution from heuristic methods like BLEU to advanced models like TransQuest and COMET. Transformer-based QE models outperformed traditional ones in capturing semantic and contextual information. While LLMs (e.g., LLaMA, Mistral, Gemma) showed potential with zero-shot and few-shot prompting, they still lagged behind fine-tuned models, especially for low-resource languages. Few-shot prompting improved LLM performance significantly but wasn't consistent across settings. TransQuest and COMET remained the most reliable due to task-specific fine-tuning. Limitations included lack of QE data for low-resource pairs and high computational cost of transformer and LLM models. Future work should focus on expanding low-resource datasets, exploring hybrid QE

architectures, and enhancing prompt design. Domain-specific fine-tuning and lightweight QE models are also recommended. Lastly, improving QE model interpretability could benefit real-world applications.

References

Alva-Manchego, F. *et al.* (2021a) 'deepQuest-py: Large and Distilled Models for Quality Estimation', *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 382–389. Available at: https://doi.org/10.18653/v1/2021.emnlp-demo.42.

Alva-Manchego, F. *et al.* (2021b) 'deepQuest-py: Large and Distilled Models for Quality Estimation', *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 382–389. Available at: https://doi.org/10.18653/v1/2021.emnlp-demo.42.

Bhattacharyya, P. *et al.* (2023a) 'Findings of the WMT 2023 Shared Task on Automatic Post-Editing', *Conference on Machine Translation - Proceedings*, (1), pp. 670–679. Available at: https://doi.org/10.18653/v1/2023.wmt-1.55.

Bhattacharyya, P. *et al.* (2023b) 'Findings of the WMT 2023 Shared Task on Automatic Post-Editing', *Conference on Machine Translation - Proceedings*, (1), pp. 670–679. Available at: https://doi.org/10.18653/v1/2023.wmt-1.55.

Chimoto, E.A. and Bassett, B.A. (2022) 'COMET-QE and Active Learning for Low-Resource Machine Translation', *Findings of the Association for Computational Linguistics: EMNLP 2022*, (2019), pp. 4764–4769. Available at: https://doi.org/10.18653/v1/2022.findings-emnlp.348.

Conneau, A. *et al.* (2020) 'Unsupervised cross-lingual representation learning at scale', *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451. Available at: https://doi.org/10.18653/v1/2020.acl-main.747.

Deoghare, S., Kanojia, D. and Ranasinghe, T. (2023) 'Quality Estimation-Assisted Automatic Post-Editing', pp. 1686–1698.

Devlin, J. *et al.* (2018) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', *Naacl-Hlt 2019*, (Mlm), pp. 4171–4186. Available at: https://aclanthology.org/N19-1423.pdf.

Dubey, A. *et al.* (2024) 'The Llama 3 Herd of Models', pp. 1–92. Available at: http://arxiv.org/abs/2407.21783.

Fernandes, P. *et al.* (2023) 'The Devil is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation', *Conference on Machine Translation - Proceedings*, pp. 1064–1081. Available at: https://doi.org/10.18653/v1/2023.wmt-1.100.

Gemma Team *et al.* (2024) 'Gemma: Open Models Based on Gemini Research and Technology', pp. 1–16. Available at: http://arxiv.org/abs/2403.08295.

Graham, Y. *et al.* (2020) 'Continuous measurement scales in human evaluation of machine translation', *LAW 2013 and ID 2013 - 7th Linguistic Annotation Workshop and Interoperability with Discourse, Proceedings of the Workshop*, pp. 33–41.

Heafield, K., Zhu, Q. and Grundkiewicz, R. (2021a) 'Findings of the WMT 2021 Shared Task on Efficient Translation', *WMT 2021 - 6th Conference on Machine Translation, Proceedings*, pp. 639–651.

Heafield, K., Zhu, Q. and Grundkiewicz, R. (2021b) 'Findings of the WMT 2021 Shared Task on Efficient Translation', *WMT 2021 - 6th Conference on Machine Translation, Proceedings*, pp. 639–651.

Iyer, V. *et al.* (2024) 'Quality or Quantity? On Data Scale and Diversity in Adapting Large Language Models for Low-Resource Translation'. Available at: http://arxiv.org/abs/2408.12780.
Jiang, A.Q. *et al.* (2023) 'Mistral 7B', pp. 1–9. Available at: http://arxiv.org/abs/2310.06825.

Kanojia, D. *et al.* (2021) 'Pushing the Right Buttons: Adversarial Evaluation of Quality Estimation', *WMT 2021 - 6th Conference on Machine Translation, Proceedings*, pp. 625–638.

Kepler, F. *et al.* (2019) 'OpenKiwi: An open source framework for quality estimation', *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations*, pp. 117–122. Available at: https://doi.org/10.18653/v1/p19-3020.

Kocmi, T. and Federmann, C. (2023) 'Large Language Models Are State-of-the-Art Evaluators of Translation Quality', *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023*, pp. 193–203.

Papineni, K. *et al.* (2002) 'BLEU: A method for automatic evaluation of machine translation', *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2002-July(July), pp. 311–318.

Qian, S. *et al.* (2024) 'A Multi-task Learning Framework for Evaluating Machine Translation of Emotion-loaded User-generated Content', *Conference on Machine Translation - Proceedings*, 2024-Novem, pp. 1140–1154.

Ranasinghe, T., Orăsan, C. and Mitkov, R. (2020) 'TransQuest: Translation Quality Estimation with Cross-lingual Transformers', *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, pp. 5070–5081. Available at: https://doi.org/10.18653/v1/2020.coling-main.445.

Ranasinghe, T., Orăsan, C. and Mitkov, R. (2021) 'An Exploratory Analysis of MultilingualWord-Level Quality Estimation with Cross-Lingual Transformers', *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2, pp. 434–440. Available at: https://doi.org/10.18653/v1/2021.acl-short.55.

Sellam, T., Das, D. and Parikh, A.P. (2020) 'BLEURT: Learning robust metrics for text generation', *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892. Available at: https://doi.org/10.18653/v1/2020.acl-main.704.

Specia, L. *et al.* (2020) 'Findings of the WMT 2020 Shared Task on Quality Estimation', (c), pp. 743–764.

Vaswani, A. *et al.* (2017) 'Attention is all you need', *Advances in Neural Information Processing Systems*, 2017-Decem(Nips), pp. 5999–6009.

Zerva, C. *et al.* (2022a) 'Findings of the WMT 2022 Shared Task on Quality Estimation', *Conference on Machine Translation - Proceedings*, (2016), pp. 69–99.

Zerva, C. *et al.* (2022b) 'Findings of the WMT 2022 Shared Task on Quality Estimation', *Conference on Machine Translation - Proceedings*, (2016), pp. 69–99.

Zhang, B. *et al.* (2024) 'Benchmarking Large Language Models for Cervical Spondylosis', *JMIR Formative Research*, 8, p. e55577. Available at: https://doi.org/10.2196/55577.

Zhu, Y. *et al.* (2023) 'Large Language Models for Information Retrieval: A Survey', pp. 1–35. Available at: http://arxiv.org/abs/2308.07107.

APPENDIX A

RESEARCH PROPOSAL

GitHub Address for Research Proposal:
Link :
https://github.com/SatyaIyerRaghav/LJMU_Thesis_Reports/tree/main/Research_Proposal

APPENDIX B:

MODEL EVALUATION SUMMARY TABLE

- **Summary of Model Performance Metrics [EN-HI]**

	GENERAL MODEL BUILDING [EN-HI]					
Sn.No	**Model Name**	**Model Family**	**Model Type**	**Spearson**	**Pearson**	**MAE**
1	Transquest/monotransquest_da_multilingual	TRANSQUEST	Xlmroberta	**0.5502**	**0.6241**	**89.5147**
2	wmt21_comet_qe_da	COMET-KIWI	BART or T5	**0.3188**	**0.3015**	**80.6831**
3	Meta-LLaMA-3.2-3B Instruct	Llama	LLaMA-3.2-3B Instruct	**0.0664**	**0.0767**	**80.1322**
4	Meta-LLaMA-3- 8B Instruct	Llama	LLaMA-3- 8B Instruct	**0.1099**	**0.1297**	**80.0512**
	Zero Shot Prompt Models [EN-HI][Encoder Models]					
Sn.No	**Model Name**	**Model Family**	**Model Type**	**Spearson**	**Pearson**	**MAE**
1	Meta-LLaMA-3.2-3B Instruct	Llama	LLaMA-3.2-3B Instruct	-0.0636	-0.0431	80.5913
2	google/gemma-1.1-4b-it	Google	gemma-1.1-4b-it	**0.0114**	**0.0072**	**11.9853**
3	mistralai/Mistral-7B-Instruct-v0.2	Mistral AI	Mistral-7B-Instruct-v0.2	-0.017	-0.0226	11.9253
	Few Shot Prompt Models [EN-HI][Decoder Models]					
Sn.No	**Model Name**	**Model Family**	**Model Type**	**Spearson**	**Pearson**	**MAE**
1	Meta-LLaMA-3.2-3B Instruct	Llama	LLaMA-3.2-3B Instruct	0.0385	0.0776	11.3172
2	google/gemma-1.1-4b-it	Google	gemma-1.1-4b-it	-0.0166	0.0005	12.0147
3	mistralai/Mistral-7B-Instruct-v0.2	Mistral AI	Mistral-7B-Instruct-v0.2	-0.0122	-0.0129	11.7673

- **Summary of Model Performance Metrics [EN-MR]**

	GENERAL MODEL BUILDING [EN-MR]					
Sn.No	**Model Name**	**Model Family**	**Model Type**	**Spearson**	**Pearson**	**MAE**
1	Transquest/monotransquest_da_multilingual	TRANSQUEST	Xlmroberta	**0.1962**	**0.1997**	**7.781**
2	wmt21_comet_qe_da	COMET-KIWI	BART or T5	**0.3882**	**0.4716**	**69.6762**
3	Meta-LLaMA-3.2-3B Instruct	Llama	LLaMA-3.2-3B Instruct	**0.0274**	**0.0848**	**69.2245**
4	Meta-LLaMA-3- 8B Instruct	Llama	LLaMA-3- 8B Instruct	**0.1283**	**0.1877**	**69.1203**
	Zero Shot Prompt Models [EN-MR]					
Sn.No	**Model Name**	**Model Family**	**Model Type**	**Spearson**	**Pearson**	**MAE**
1	Meta-LLaMA-3.2-3B Instruct	Llama	LLaMA-3.2-3B Instruct	0.052	0.0415	14.9053
2	google/gemma-1.1-4b-it	Google	gemma-1.1-4b-it	**0.0313**	**0.0362**	**15.1698**
3	mistralai/Mistral-7B-Instruct-v0.2	Mistral AI	Mistral-7B-Instruct-v0.2	0.045	0.0573	14.8112
	Few Shot Prompt Models [EN-MR]					
Sn.No	**Model Name**	**Model Family**	**Model Type**	**Spearson**	**Pearson**	**MAE**
1	Meta-LLaMA-3.2-3B Instruct	Llama	LLaMA-3.2-3B Instruct	0.0509	0.0434	14.8123
2	google/gemma-1.1-4b-it	Google	gemma-1.1-4b-it	0.0387	0.0571	14.7093
3	mistralai/Mistral-7B-Instruct-v0.2	Mistral AI	Mistral-7B-Instruct-v0.2	0.1087	0.1104	14.2438

APPENDIX-C

PYTHON SOURCE CODE ,DATASET , VIDEO LINKS

GitHub Address for source code
Link: https://github.com/SatyaIyerRaghav/LJMU_Thesis_Reports/tree/main

GitHub Address for Dataset
Link: https://github.com/SatyaIyerRaghav/LJMU_Thesis_Reports/tree/main/DataSet

GitHub Address for Video
Link:
https://github.com/SatyaIyerRaghav/LJMU_Thesis_Reports/tree/main/Video_Presentation

MT QE Thesis Report By Satya Iyer.docx

ORIGINALITY REPORT

4%	3%	1%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Liverpool John Moores University Student Paper	1%
2	aclanthology.org Internet Source	1%
3	www.arxiv-vanity.com Internet Source	1%
4	www.statmt.org Internet Source	1%
5	arxiv.org Internet Source	1%

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	On		