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

Presenter **Satya Padmanabhan Iyer**

University: Liverpool John Moores University, UK | Date: May 2025

Agenda

1 Introduction

Recent advances in Large Language Models (LLMs) have significantly improved Machine Translation (MT). However, traditional evaluation metrics fall short without reference translations. Quality Estimation (QE) offers a reference-free solution by predicting translation quality directly.

Here are the key points extracted from the provided Thesis introduction:

1. **Limitations of Traditional MT Evaluation**: Conventional metrics like BLEU, BLEURT, and BERTScore rely on reference translations, which are often unavailable in real-world scenarios. These metrics also fail to effectively capture nuances such as idioms, paraphrasing, and domain-specific language.
2. **Emergence of Quality Estimation (QE):** QE methods estimate the quality of machine translations without reference translations by using models fine-tuned on human-annotated data like Direct Assessment (DA) scores. This allows for more practical and nuanced evaluation.
3. **Prominent QE Models – TransQuest and COMET:**
    1. *TransQuest* is efficient and multilingual, particularly strong in low-resource settings.
    2. *COMET* excels in semantic and fluency prediction by leveraging DA scores and multilingual encoders.

    These models offer complementary strengths in MT quality estimation.
4. **Challenges with Public LLMs in Multilingual Settings:** Many open-source LLMs are optimized for English and underperform in multilingual tasks compared to proprietary models like GPT-4. This highlights the need for fine-tuning and adaptation for multilingual applications.
5. **Contributions of my Research Paper:**
    - Proposes a simple architecture for sentence-level QE.
    - Demonstrates that multilingual QE models are competitive with bilingual ones.
    - Explores zero-shot and few-shot transfer capabilities across domains, MT types, and language pairs.
    - Releases code and models as part of an open-source framework.

2 Literature Review

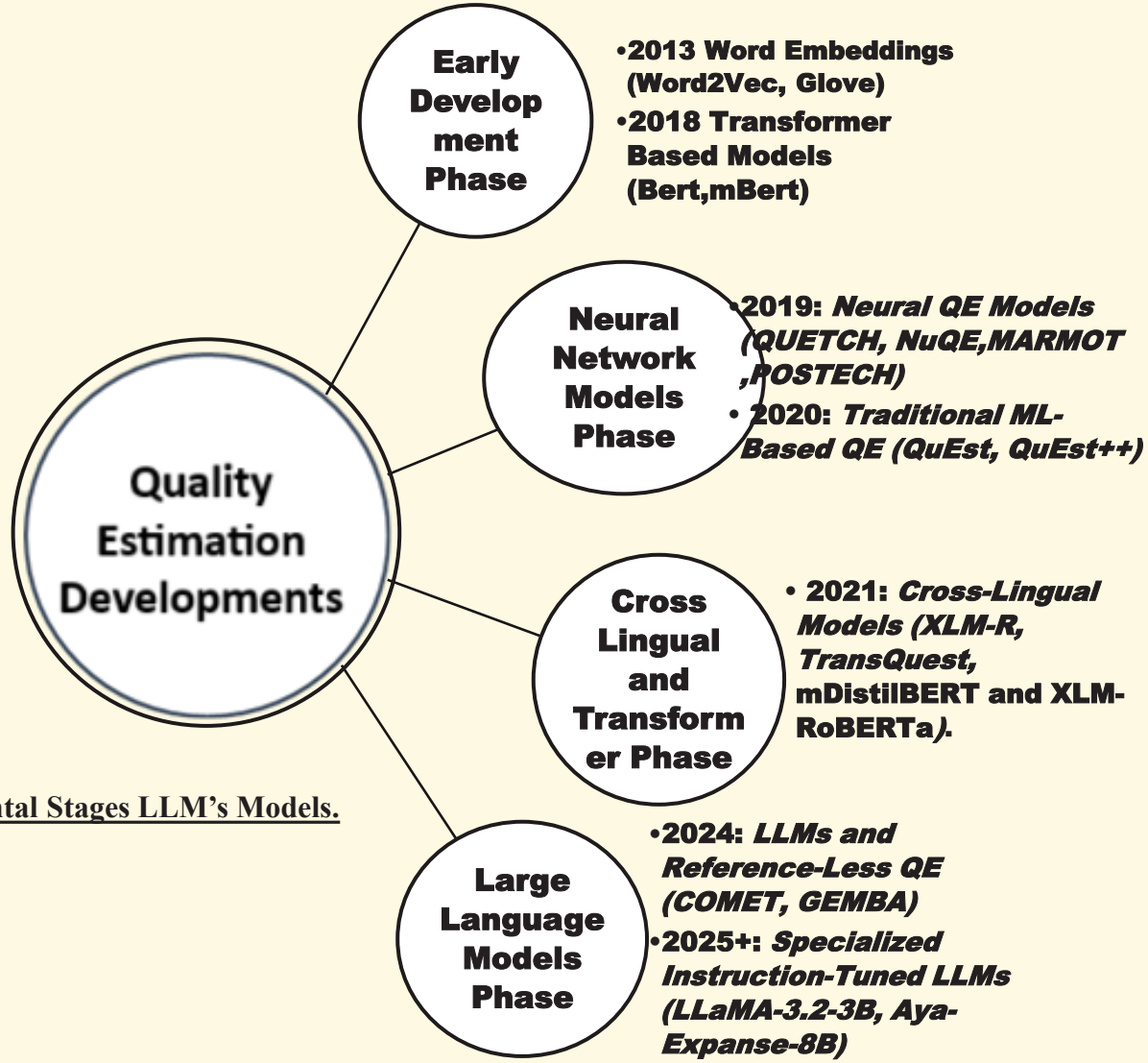

Figure 2: Detailed Evolution and Developmental Stages LLM's Models.

2 Literature Review (contd..)

Key Challenges in MT Quality Estimation (QE)

- **High Dependence on Human Annotations**
  Existing models (TransQuest, COMET) require costly Direct Assessment (DA) scores, limiting scalability to low-resource pairs (e.g., en-hi, en–mr).

- **Multilingual Model Limitations**
  Models like XLM-R struggle with low-resource languages due to the "curse of multilingualism."

- **Semantic Weaknesses**
  Current QE systems lack deep semantic reasoning for idioms, ambiguity, and domain-specific terms.

- **Ineffective Metrics in Zero-Reference Scenarios**
  Lexical metrics (BLEU, TER) fail to capture fluency and adequacy without reference translations.

- **Underutilization of Instruction-Tuned LLMs**
  LLMs like LlaMA-3, Mistral, and Gemma remain underexplored for QE in few-shot/zero-shot contexts.

Proposed Framework & Contributions

Instruction-Tuned LLMs for Low-Resource QE

•LLM-based QE Framework
Uses LlaMA-3.2-3B-Instruct, Mistral-7B-Instruct, Gemma-1.1-4B-IT with zero-shot/few-shot prompting.

•Low-Resource Focus
Tailored for English–Hindi (en–hi) and English–Marathi (en–mr) using WMT datasets.

•Reduces Human Annotation Dependence
Enables semantic evaluation without reference translations.

•Leverages Instruction Tuning & Cross-Lingual Transfer
Improves contextual understanding and generalization.

•Benchmarking Against SOTA
Compared against TransQuest and COMET using DA, BLEU, BERTScore, and COMET.

3 Problem Statement

Challenge in QE without References: Traditional Machine Translation (MT) Quality Estimation (QE) models depend heavily on large multilingual encoders (e.g., XLM-R), but struggle in generalizing across domains, languages, and linguistic diversity—especially when no reference translations are available.

Multilingual Adaptation Difficulty: Traditional QE models face limitations in handling idiomatic expressions, domain-specific terms, and highly diverse sentence structures.

Our Proposed Approach:
Introduces **Zero-Shot Learning (ZSL)** and **Few-Shot Learning (FSL)** for sentence-level QE. Enables **training once, deploying across many languages** without requiring labeled examples for each. Significantly improves **scalability and adaptability** for real-world MT systems.

Zero-Shot Translation Use Case:
Accurately interprets ambiguous terms (e.g., "bank") in Hindi and Marathi **without prior language-specific training**, validating the model's contextual reasoning capabilities.

Few-Shot Translation Use Case:
With just 1–2 examples, the model translates new sentences effectively, achieving high QE scores (>85%)—demonstrating fast domain adaptation.

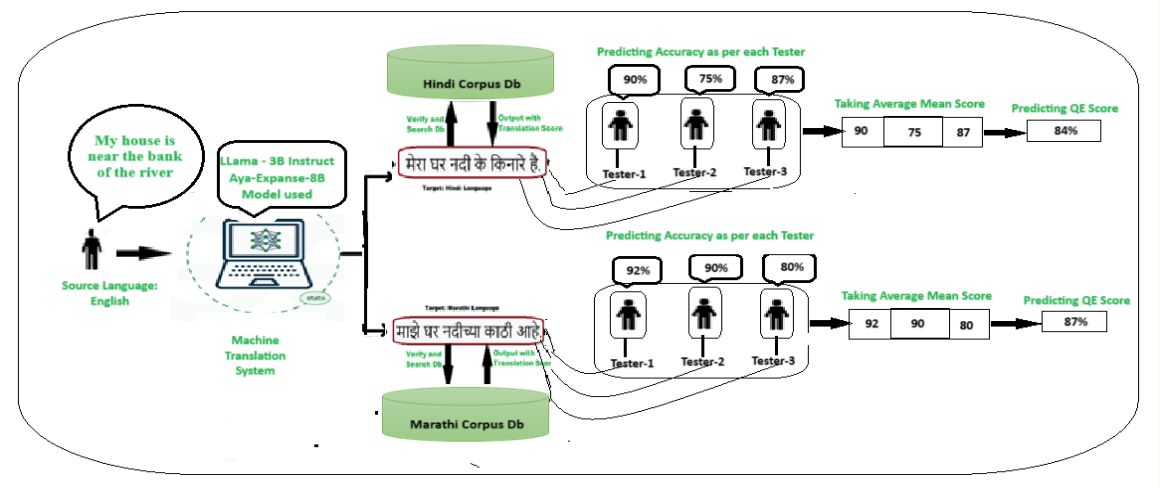

Figure 1: Architecture Design of Machine Translation with Quality Estimation Scores from English-Hindi and English-Marathi

QE Scoring System:
Human evaluators assign sentence-level scores → Mean QE score calculated → Final Quality Prediction produced (e.g., 87% for Hindi, 86% for Marathi).

Our Contributions:

- Benchmark LLMs using prompt-based QE against fine-tuned PTLMs.
- Analyze the importance of **prompt design** and **reference inclusion**.
- Reveal that LLMs generalize well but need optimization for domain-specific QE.

4 Methodology

Datasets :
•**WMT23 QE Shared Task** via Hugging Face
(English–Hindi & English–Marathi pairs with DA scores)

•**train/dev.tsv files**:
- En–Hi: 7K train / 1K dev | Mean ≈ 80
- En–Mr: 26K train / 1K dev | Mean ≈ 70, wider variance

Preprocessing: Tokenizers:
- **TransQuest**: XLM-RoBERTa
- **COMET**: SentencePiece (XLM-R-Large)
- **LLaMA, Mistral, Gemma**: Instruction-tuned BPE / ChatML formats

Models Used:
•LLaMA-3.2-3B-Instruct
•Mistral-7B-Instruct-v0.2
•Gemma-1.1-4B-IT

Tasks:
•Sentence-level QE
•Generative QE
•Justification analysis (qualitative)

Evaluation Metrics:
•Pearson/Spearman Correlation (score alignment)
•MAE / RMSE (error)
•Inference speed & memory (efficiency)
•Explanation quality (user trust)

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

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

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

4i Methodology (Contd..)

Step 1: Problem Formulation

QE → Classification (instead of regression):

- **0**: Poor (0.0–0.3)
- **1**: Moderate (0.3–0.7)
- **2**: High (0.7–1.0)

Easier to interpret & integrate (e.g., re-ranking, filtering)

Step 2: Input Prompt Format

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:

During preprocessing, Human Direct Assessment (DA) scores were mapped to the appropriate class labels (0, 1, or 2). 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.

Methodology (Contd..)

Tools & Technologies for MT Quality Estimation (QE)

Programming Environment

•**Python 3.10+** – Core scripting for data, modeling, and evaluation

•**Google Colab** – Dev environment with free **GPU (Tesla T4)** for fine-tuning & inference

Key Libraries & Frameworks

•**NumPy, Pandas** – Data handling & metric calculations

•**Scikit-learn** – Evaluation: Pearson, Spearman, MAE, RMSE

•**Plotly** – Interactive visualizations

NLP & Deep Learning

•**Hugging Face Transformers** – LLaMA-3, Mistral-7B, Gemma-4B for prompt-based QE

•**SentenceTransformers** – Baselines with LaBSE, MPNet, etc.

•**TransQuest, COMET** – Used for comparative benchmarking

Prompt Engineering

•Crafted task-specific prompts for LLMs to classify translation quality with justification

5 Result and Discussion

1. Train vs Dev QE Score Correlation (Observations):

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.

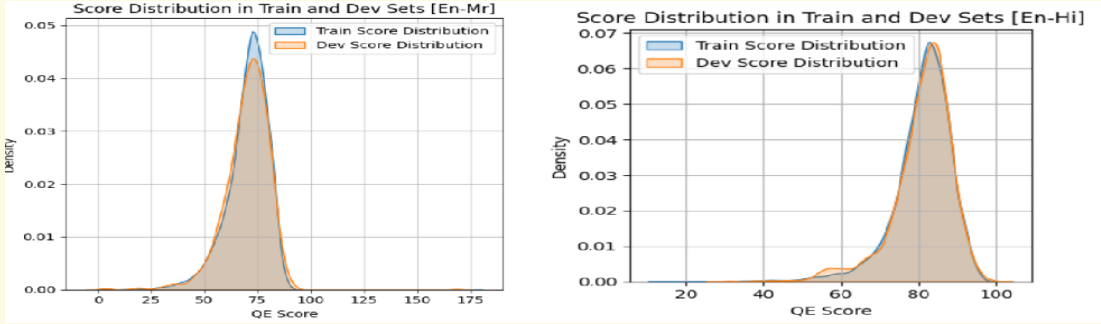

Figure: 4 Statistical Study - Score Distribution [En-hi] and [En-Mr]

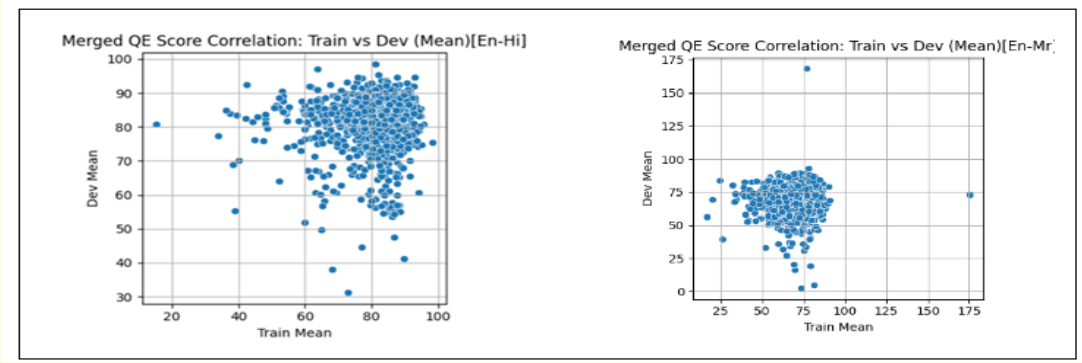

Figure 10: Train vs. Dev Score Correlation [En-hi] and [En-Mr]

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.

5i Result and Discussion(Contd..)

2. Predicted Score vs Human Mean Score (Observation)

- **TransQuest & COMET**: Show strong linear correlation with human scores due to task-specific training.
- **Zero-shot LLMs (e.g., LLaMA, Mistral)**: Exhibit wider score spread due to lack of QE-specific tuning.
- **Few-shot Prompting**: Significantly reduces error spread, improving alignment with human judgment.

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

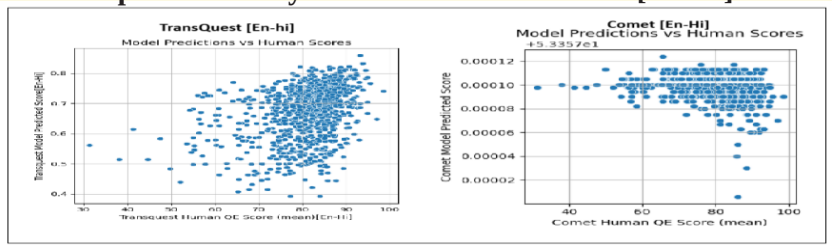

Figure 11: TransQuest and Comet Model [En-Hi]

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

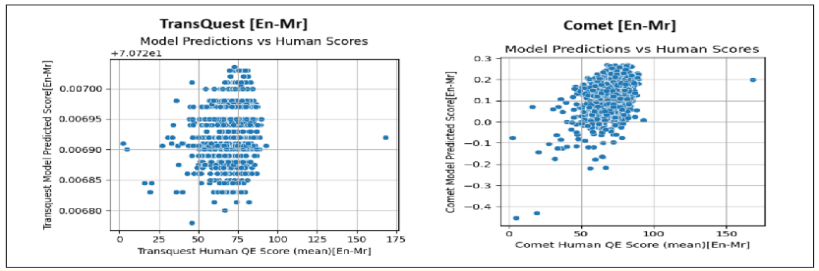

Figure 13: TransQuest and Comet Model [En-Mr]

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

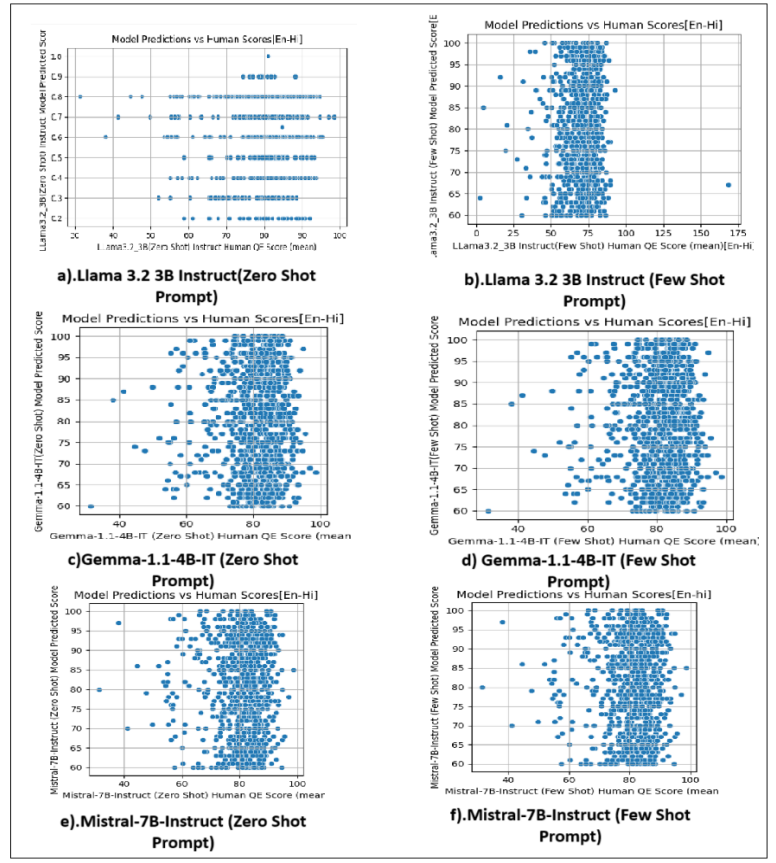

Figure 12: Compares model-predicted QE scores vs human QE scores (mean values) across three different models (in both zero-shot and few-shot prompt settings for [En-Hi].

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

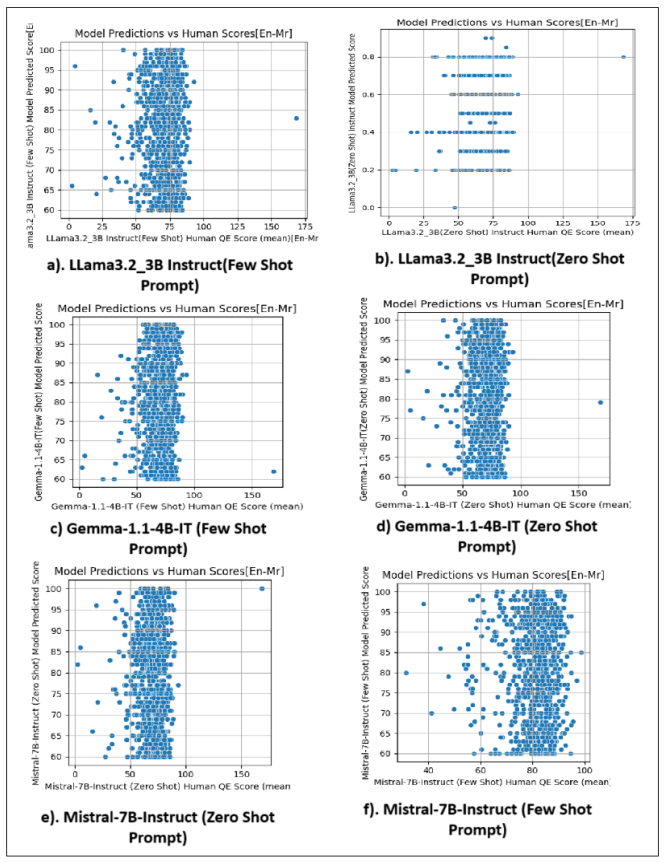

Figure 14: Compares model-predicted QE scores vs human QE scores (mean values) across three different models in both zero-shot and few-shot prompt settings for [En-Mr].

5ii Result and Discussion(Contd..)

3. Error Score Distribution Analysis (Observation):

•**En-Hi**: COMET outperforms TransQuest with lower and tighter error distribution.
•**LLMs (Few-shot vs Zero-shot)**: Few-shot prompting consistently reduces errors, especially in Mistral-7 B.
•**En-Mr**: TransQuest outperforms COMET; again, few-shot prompting improves all LLMs' accuracy.

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

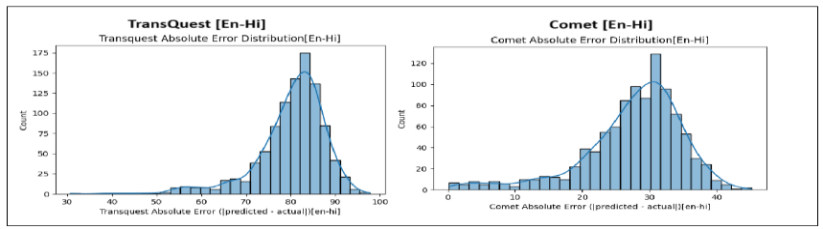

Figure 15: Correlation and Error Metrics for [En-Hi]

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

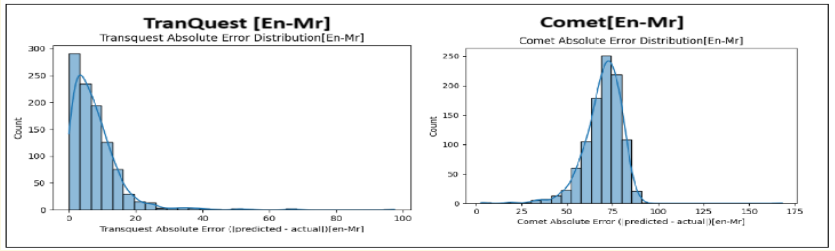

Figure 17: Correlation and Error Metrics of [En-Mr]

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

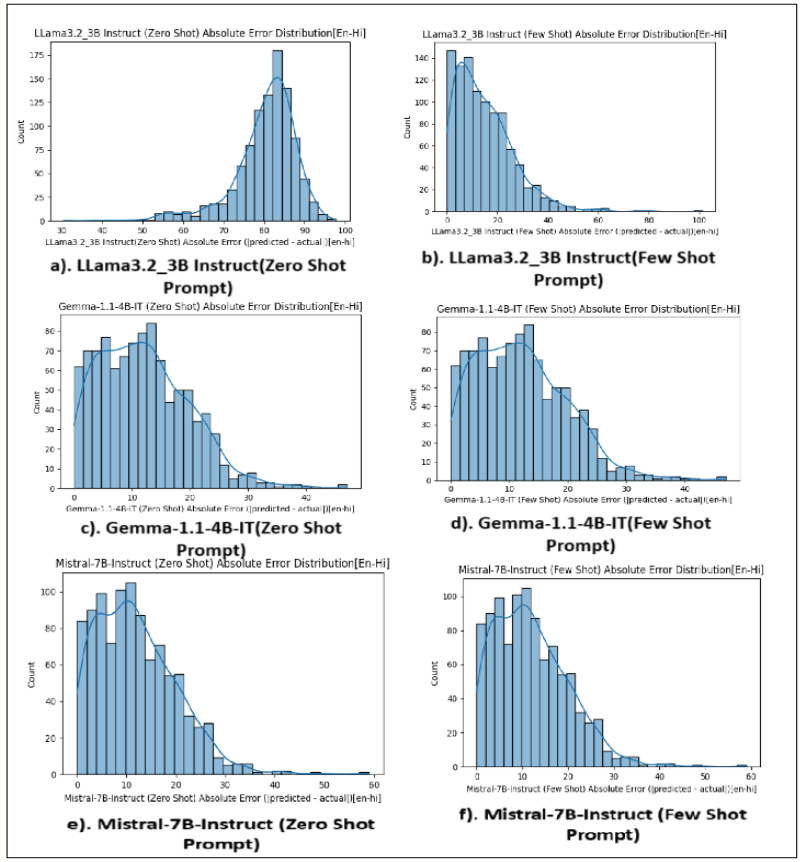

Figure 16 : Correlation and Error Metrics across three different models in both zero-shot and few-shot prompt settings for [En-Hi].

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

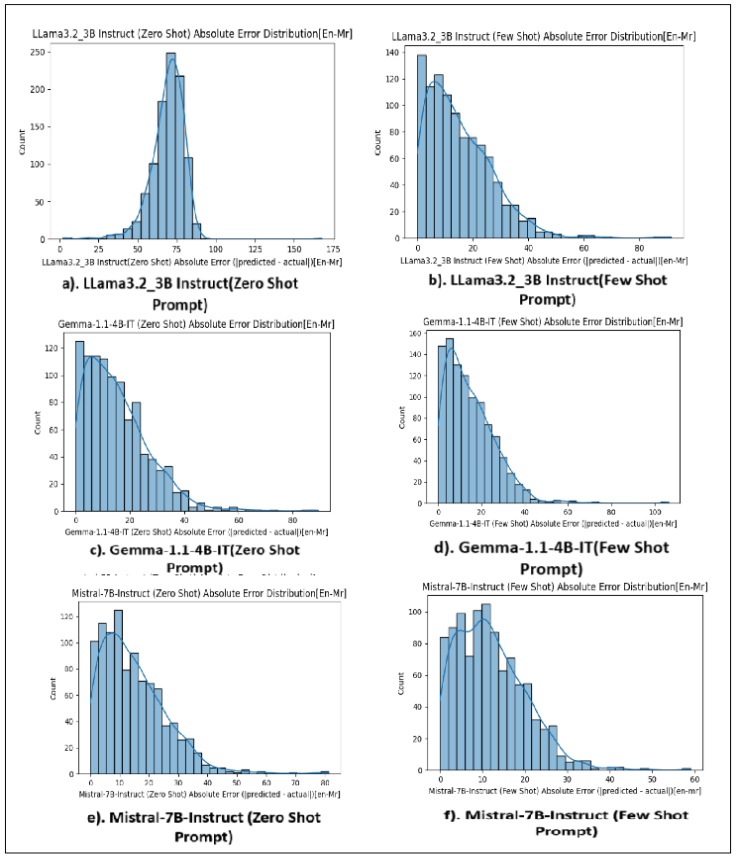

Figure 18: Correlation and Error Metrics across three different models) in both zero-shot and few-shot prompt settings for [En-Mr].

5iii Result and Discussion(Contd..)

4. Evaluation of Sampling Methods and Results

Three primary sampling strategies were evaluated:

1. Zero-shot prompting: No training examples were provided. LLMs performed inconsistently with higher variance in predictions.

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

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:
"""
```

Conclusion from Zero-Shot Results (Tables 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

5iv Result and Discussion(Contd..)

2. Few-shot prompting:

A Few examples were provided for referencing.

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

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.

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).

Conclusion from Few Shot (Tables 9 &10):

From Tables 9 and 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. Meta-LLaMA-3.2-3B showed slightly better correlations for English-Hindi, but still inadequate. Mistral-7B outperformed others in both Spearman and Pearson for English-Marathi, 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

5v Result and Discussion(Contd..)

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.

Conclusions for Fine Tuning (Tables 11 &12):

From **Tables 11 and 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

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

6 Conclusion and Future Work

6.1 Evolution and Advancements in Machine Translation Quality Estimation Models

Shifted from reference-based metrics (BLEU, METEOR, TER) to reference-free, deep learning-based methods. Early models lacked semantic and syntactic understanding. Transformer-based models (BERT, mBERT, XLM, XLM-R) enabled multilingual, context-aware QE. TransQuest and COMET improved sentence-level prediction accuracy. Research evaluated both traditional metrics and advanced transformer-based and LLM models (LLaMA, Mistral, Gemma), showing the evolution toward intelligent and multilingual QE systems.

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

Limitations of traditional metrics: Focus on surface-level similarity (token overlap, edit distance).Poor at semantic, fluency, and contextual evaluation.
Advantages of transformer-based QE models: Use pretrained embeddings for semantic understanding. Fine-tuned for QE tasks to improve performance. Support multilingual settings effectively.
Challenges:

- Need for labelled QE datasets.
- High computational requirements.
- Domain generalization issues.

Conclusion: Transformer-based models outperform traditional metrics, especially in complex, multilingual settings.

6.3 Evaluating Large Language Models for Enhanced Translation Quality Estimation

Use of instruction-tuned LLMs with zero-shot and few-shot prompting.
Zero-shot prompting: Mixed results, high variance, inconsistent human correlation.
Few-shot prompting: Improved results (LLaMA, Mistral), competitive with TransQuest and COMET.
Limitations:

- LLMs are less effective in low-resource languages due to data scarcity.
- LLMs are decoder-based (generation-focused), whereas QE tasks benefit from encoder-based models (like TransQuest, COMET).

Conclusion: LLMs offer flexible alternatives in low-resource or zero-data scenarios, but dedicated QE models remain superior for high-accuracy tasks.

6.3 FUTURE RECOMMENDATIONS

•Develop high-quality QE datasets for low-resource language pairs.
•Explore hybrid QE models combining encoder-based and decoder-based architectures.
•Improve prompt engineering for LLM-based QE.
•Fine-tune QE models on domain-specific data (e.g., medical, legal).
•Create lightweight QE models for real-time or mobile applications.
•Enhance model interpretability, especially for LLMs, to build trust and diagnose translation issues.

7 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

8 Source-Code Link

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

Thank you

Satya Padmanabhan Iyer