**APPENDIX A**

**RESEARCH PROPOSAL**

**Abstract**

Machine Translation evaluation using Quality Estimation methods traditionally requires encoder-based models to perform a regression task that predicts the translation quality score on a scale of 0–100. We use existing WMT data sets that contain key information like Source Sentences, Machine Translation output, and Reference translations. These are crucial for MT quality evaluation, while translation errors are less significant. Traditional metrics use lexical comparison to evaluate the quality of MT against a single reference translation, leading to vague semantic matching and ignoring translations that can be idiomatic/metaphorical/domain-specific. Our work aims to train computational models to evaluate translation quality using Large Language Models within the GenAI paradigm to perform well.

We will compare our results with baseline approaches like TransQuest, COMET, and statistical metrics like BLEU, TER, and chRF. The proposed approach improves LLaMA-3.2-3B-Instruct, and Aya-Expanse-8B by focusing on a multilingual corpus and instruction fine-tuning high-quality translation-related instructions. The resulting model should outperform open-source alternatives and should be competitive with closed models like GPT-4, aiming to demonstrate its effectiveness in handling multiple translation tasks.

**1. Introduction**

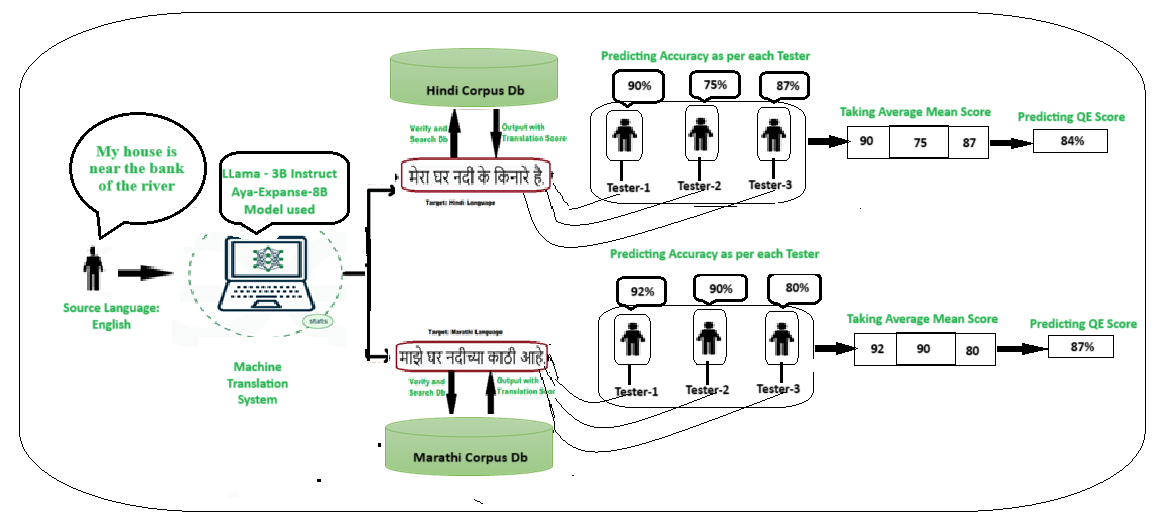
Numerous crucial responsibilities in multilingual NLP, including quality assessment, automated revision, and linguistic error correction, require examining, producing, or handling text across different languages and play a key role in diverse translation processes. Lately, versatile large-scale language models (LLMs) have challenged the conventional approach of task-specific systems, attaining cutting-edge results in multiple recent WMT-shared challenges (Kocmi & Federmann, 2023a).Regrettably, robust proficiency in various translation-associated tasks has thus far been demonstrated solely by proprietary LLMs, possibly because most publicly available LLMs prioritize English, causing methods utilizing these models to still lag, achieving competitive outcomes only when focusing on a singular objective.

In our work, we tackle this issue by creating multilingual sentence-level QE models that deliver strong results across various fields, machine translation styles, and language combinations. Furthermore, for the first time, we introduce sentence-level QE as a zero-shot cross-lingual adaptation task, opening up new research possibilities where multilingual models can be trained a single time and subsequently applied to a wide range of languages and domains.

TransQuest (Ranasinghe, Orăsan and Mitkov, 2020) is a new, efficient framework for sentence-level QE that reduces resource requirements and enhances adaptability, making it suitable for low-resource languages. TransQuest has been shown to outperform current state-of-the-art QE frameworks across 15 language pairs and has won several shared tasks at WMT 2020.

The rise of LLMs in NLP tasks, such as question answering, has significantly improved applications like machine translation (MT), text summarization, and information retrieval (Kocmi and Federmann, 2023; Zhu *et al.*, 2023) .Traditional MT quality assessment relies on metrics like BLEU, BLEURT, and BERTScore(Papineni *et al.*, 2002; Sellam, Das and Parikh, 2020; Zhang *et al.*, 2024) by comparing Machine Translation output to reference translations. When references are unavailable, QE techniques fine-tune multilingual pre-trained models on human evaluation datasets like DA scores (Graham *et al.*, 2020; Kanojia *et al.*, 2021; Zerva *et al.*, 2022a). Recent research explores prompting LLMs to assign translation quality scores, showing promising results (Kocmi and Federmann, 2023) DA scores, ranging from 0 to 100, reflect translation quality based on human ratings and are standardized into z-scores for training QE models (Graham et al., 2020).

Additionally, XLM-R, a multilingual masked language model trained on text from 100 languages, balances performance between high- and low-resource languages. The study reveals that, by increasing model capacity, the "curse of multilingualism" can be alleviated, improving cross-lingual performance. Figure 1 Explains the architectural diagram of Woking style of Machine Translation with Quality Scores.



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

Fig1 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 **Chain of Thought (CoT) reasoning** can be applied in this context.

* 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 **Chain of Thought (CoT) Translation Methodology**

It improves model outputs by breaking down the problem into intermediate logical steps before generating the final response. It is particularly useful when dealing with **ambiguity, complex structures, or long-form text.**

#### **Step 1: Input Sentence (English):** ➡ **"My house is near the bank of the river."** *(The word "bank" is ambiguous—does it refer to a financial bank or a riverbank?)*

#### **Step 2: Chain of Thought Reasoning:**

#### **"My house" →** No ambiguity, clearly refers to a home.

#### **"Bank" →** Could mean a financial institution or riverbank.

#### **"River" →** Helps determine that "bank" refers to the riverbank,

Since "bank" is followed by "river," the model **logically deduces** that "bank" must mean **"riverbank"** and not a financial bank.

1. **Sentence Structuring According to Target Language Grammar:**

* Hindi: In Hindi, the structure should be **"house" (मेरा घर) + "riverbank" (नदी के किनारे) + "is" (है)**
* Marathi: Similarly, in Marathi, **"house" (माझे घर) + "riverbank" (नदीच्या काठी) + "is" (आहे)**

1. **Generate Translations Using the Derived Meaning:**

* **Hindi Output:** "मेरा घर नदी के किनारे है।"
* **Marathi Output:** "माझे घर नदीच्या काठी आहे।"

#### **Step 3: Verification and QE Scoring:**

* Compared with corpus data, translation quality is checked.
* The scores 84% and 87% reflect accuracy per sentence, showing high correctness.

The model **explicitly reasons** through potential meanings before translating. **It avoids misinterpretation errors by understanding the context step-by-step** before producing output.  
**More effective for ambiguous sentences** compared to Zero-shot, where reasoning is implicit.

**Our contributions are:**

* We examine the translation details LLMs need for evaluation, including source text, references, errors, and annotation instructions.
* We analyze zero-shot Prompting, Chain Of Thoughts Prompting, and few-shot prompting for Machine Translation quality assessment.
* We compare our prompting techniques with fine-tuned multilingual PTLMs, finding LLMs still lag behind.
* Our analysis of prompt structures shows that reference texts are crucial for accurate translation assessment.

**2. Related Work**

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, Orǎsan 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 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 and Aya-Expanse-8B, which are LLMs used in natural language processing tasks due to their unique strengths and capabilities.

**Fig:2 Detailed Evolution and Developmental Stages LLM’s Models.**

**3. Research Questions**

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 and Aya-Expanse-8B provide more accurate and efficient translation quality estimation concerning state-of-the-art QE models and traditional evaluation metrics?

**4. Aim and Objective**

This research aims to propose results with baseline approaches such as TransQuest, COMET, and statistical metrics like BLEU, TER, and chrF, which are commonly used for machine translation evaluation. BLEU measures n-gram overlap between machine translations and references, with a brevity penalty to avoid short outputs, but struggles with semantic and grammatical nuances. TER assesses the number of edits needed to transform a machine translation into a reference, highlighting post-editing effort but sometimes penalizing stylistic variations. chrF evaluates character n-gram matches, effectively capturing subtle differences and partial matches, particularly for morphologically rich languages.

The proposed approach improves LLaMA-3.2-3B-Instruct and Aya-Expanse-8B are LLMs used in natural language processing tasks due to their unique strengths and capabilities. LLaMA-3.2-3B-Instruct, with 3 billion parameters, is optimized for instruction-following tasks, making it highly effective for applications like summarization, question-answering, and performing specific commands. On the other hand, Aya-Expanse-8B, with 8 billion parameters, offers a greater capacity to handle complex tasks, providing a deeper understanding of language patterns and context. This model excels in high-resource and intricate scenarios, such as advanced reasoning, multilingual tasks, and creative language generation.

The research objectives, aligned with this study's aim, are:

* To analyze the complementary strengths of LLaMA-3.2-3B-Instruct and Aya-Expanse-8B models in addressing diverse NLP tasks.
* To evaluate the efficiency of LLaMA-3.2-3B-Instruct in processing straightforward tasks.
* To assess the performance of Aya-Expanse-8B in handling complex and demanding use cases.
* To compare the two models to identify trade-offs between computational cost, scalability, and output quality.
* To provide insights into selecting the most appropriate model for specific NLP applications.

**5. Significance of The Study**

This study is significant for its potential to transform the evaluation of machine translation (MT) systems and enhance the quality estimation methods currently employed in the field. Traditional MT evaluation relies heavily on encoder-based models for regression tasks that predict translation quality scores, typically using lexical comparisons with reference translations. However, these approaches fail to address semantic nuances and ignore idiomatic, metaphorical, and domain-specific translations. This research seeks to close this gap by leveraging LLMs within the GenAI paradigm to improve the accuracy and effectiveness of translation quality assessments.

By utilizing existing WMT datasets, which contain crucial information such as source sentences, MT outputs, and reference translations, the study builds a comprehensive evaluation framework that incorporates traditional metrics like BLEU, TER, and chrF, alongside more advanced approaches like TransQuest and COMET. These metrics provide complementary insights into linguistic accuracy, fluency, and post-editing effort, offering a more nuanced understanding of translation quality. The integration of LLaMA-3.2-3B-Instruct and Aya-Expanse-8B models, fine-tuned on multilingual corpora with high-quality translation instructions, promises to outperform current open-source alternatives and demonstrate competitiveness with closed models like GPT-4.

The proposed approach will enable more accurate MT evaluation, particularly for handling complex, multilingual tasks and diverse translation contexts. It will also guide model selection by highlighting trade-offs between computational cost, scalability, and output quality, providing essential insights for informed decision-making in NLP applications. Ultimately, the study’s outcomes will drive advancements in the MT evaluation process, making it more adaptable to real-world challenges.

**6. Scope of the study**

The study will use existing datasets such as the WMT data, which includes a wide range of language pairs, covering both widely spoken and lesser-known languages. The multilingual nature of these datasets allows the study to explore translation tasks across different languages and domains. The study will examine various traditional machine translation evaluation metrics, including BLEU, TER, and chrF, alongside newer quality estimation approaches such as TransQuest and COMET. These metrics will be applied to evaluate MT systems using different methods of translation quality scoring. The scope of evaluation will encompass both lexical-based and semantic-based approaches.

The research will primarily focus on the performance of two LLaMA models—LLaMA-3.2-3B-Instruct and Aya-Expanse-8B. The scope includes investigating the strengths and limitations of these models, particularly in handling translation tasks that require nuanced semantic understanding and context, with a focus on instruction-following capabilities and computational efficiency. The scope includes benchmarking these models and evaluating trade-offs concerning computational cost, scalability, and translation quality.

In summary, the scope of this study will be comprehensive in terms of the range of languages, MT evaluation metrics, model types, and tasks, providing valuable insights into the effectiveness of LLaMA-based models for improving MT evaluation and the quality of translation outputs.

**7. Research Methodology**

The research methodology for this study will adopt a systematic approach to evaluate and compare the performance of LLMs in MT quality estimation. This study will utilize existing datasets The primary datasets used will include:

WMT23 Dataset available at Hugging Face WMT23, this dataset contains source sentences, machine translation outputs, and reference translations, which will serve as a key foundation for the evaluation.Llama-2-QE-2023-EnHi-Test available at Hugging Face Llama-2-QE-2023-EnHi-Test, this dataset is designed specifically for quality estimation and includes English-to-Hindi translation pairs.Llama-2-QE-2023-EnMr-Test available at [Hugging Face Llama-2-QE-2023-EnMr-Test](https://huggingface.co/datasets/dipteshkanojia/llama-2-qe-2023-enmr-test), this dataset covers English-to-Marathi translations for quality estimation purposes.

LLaMA-3.2-3B-Instruct, optimized for instruction-following tasks, will be tested for straightforward translation tasks, while Aya-Expanse-8B, which handles more complex tasks, will be evaluated for multilingual and intricate translation scenarios. Both models will be implemented using pre-trained versions and fine-tuned on a multilingual corpus with high-quality translation-related instructions. Evaluation will be done on translation tasks, including multilingual translations, summarization, and creative language generation. The LLaMA models will be compared to closed models, such as GPT-4, to determine how they perform relative to proprietary models in translation tasks.

The final stage of the research methodology involves the analysis and interpretation of results. This will include identifying the trade-offs between model performance, computational cost, and scalability. Insights gained from the model evaluations will inform recommendations for selecting the most appropriate model for specific translation applications

**8.Required Resources**

The following resources will be required for the successful completion of this research:

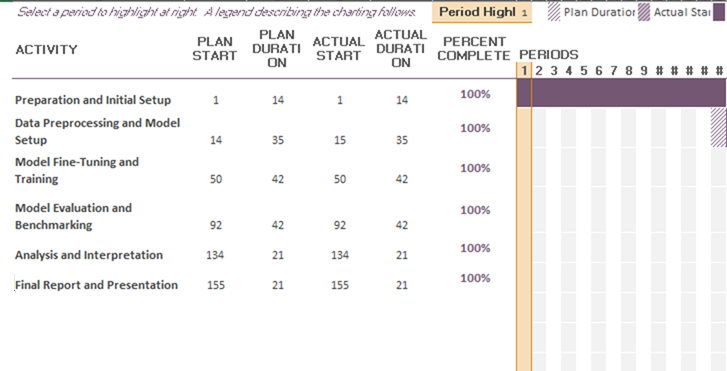
**Hardware Requirement:** To handle the large-scale processing demands of training and fine-tuning LLMs, access to HPC systems or cloud-based GPU/TPU clusters is necessary. Google TPUs will be essential for model training, fine-tuning, and evaluation, especially when working with large models like LLaMA-3.2-3B-Instruct and Aya-Expanse-8B.

**Storage:** Sufficient storage for dataset management and model checkpoints, particularly for the large WMT datasets and fine-tuned models, is needed. Cloud storage solutions or local high-capacity storage devices will be required.

**Software Requirement**: Python will be the primary language for data processing, model training, and evaluation. Libraries such as PyTorch, TensorFlow, Hugging Face Transformers, and Hugging Face Datasets will be essential for model training, fine-tuning, and performance evaluation.to pre-trained models from repositories such as Hugging Face Model Hub (e.g., LLaMA-3.2-3B-Instruct and Aya-Expanse-8B) will be needed for fine-tuning and comparison purposes. Tools for calculating evaluation metrics like BLEU, TER, and chrF, and quality estimation tools like COMET and TransQuest will be required to assess the models' performance.

**9. Research Plan**

The research plan outlines the key activities, timelines, and milestones for studying machine translation quality estimation using LLMs. The plan is structured in distinct phases, ensuring that each aspect of the research is thoroughly addressed.



By following the research plan discussed above, the study will ensure systematic progress, timely completion, and a thorough evaluation of the proposed approach for machine translation quality estimation using Large Language Models.