Master Thesis

Knowledge Distillation for Machine Translation Quality Estimation

Ningxuan Guo

Supervisor	Sophie Arnoult
2^{nd} reader	Pia Sommerauer

a thesis submitted in fulfillment of the requirements for the degree of

MA Linguistics

(Text Mining)

Vrije Universiteit Amsterdam

Computational Linguistics and Text-Mining Lab
Department of Language and Communication
Faculty of Humanities

Date	August 15, 2025
Student number	2818054
Word count	16624

Abstract

This thesis investigates reasoning-enhanced knowledge distillation for Machine Translation Quality Estimation (MTQE). The goal is to distill the capabilities of a small open-source language model (Qwen3-0.6B) from two complementary teachers: human annotators providing gold Zmean scores and GPT-4o-mini generating natural language rationales. Experiments use WMT 2021–2023 Direct Assessment datasets across 17 language pairs. Two student models are compared: one fine-tuned solely on gold scores and another trained with concatenated GPT reasoning. Results show that reasoning-enhanced distillation improves Pearson correlation overall and in extreme quality ranges and long sentences, while gold-score-only training achieves lower absolute errors and better calibration in common quality ranges. Error analysis reveals that GPT reasoning increases ranking sensitivity but may introduce variance in magnitude estimation due to stylistic or evaluative biases. The study demonstrates that combining human and LLM supervision can yield complementary benefits, and discusses strategies for balancing their contributions. Reproducibility is ensured through open datasets, model releases, and detailed training specifications, with future work focusing on selective reasoning augmentation and language-specific optimization.

The code and datasets used in this work are publicly available at `https://github.com/Ningkwoknx/Distill_with_llm`.

Declaration of Authorship

I, author, declare that this thesis, titled *Knowledge Distillation for Machine Translation Quality Estimation* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: <August 15, 2025>

Signed: <Ningxuan Guo>

Acknowledgments

I would like to express my sincere gratitude to Amir Soleimani, my second supervisor during my internship at TAUS, for his guidance, encouragement, and constructive feedback throughout the project. His support and timely advice greatly contributed to keeping me on track and improving the quality of my work.

I would like to thank Piek Vossen, Antske Fokkens, Sophie Arnoult, Pia Sommerauer, and Luis da Costa. Without their help and support, it would not have been possible for me to persevere and complete this thesis.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Machine Translation (MT) has reached unprecedented quality levels with the usage of pretrained large language models(LLM) and advanced evaluation metrics. However, it remains a challenge to measure the translation qualities, especially in real world applications where human judgements are not always available. Machine Translation Quality Estimation (MTQE) addresses this challenge by predicting the quality of translations directly from the source and target sentences. While state-of-the-art MTQE systems achieve strong correlation with human judgments, they often rely on computationally expensive models that are impractical for many real-world scenarios. This thesis explores whether recent advances in Knowledge Distillation (KD) and Large Language Model (LLM)-based annotation can be combined to produce smaller, efficient, and accurate QE models—bridging the gap between high performance and deployability.

1.1 Knowledge Distillation

Knowledge Distillation (KD) is a technique to compress a model, where a compact student model learns to mimic the behavior of a usually larger model with high-performance(Hinton et al., 2015). Instead of training on hard labels only, the student learn from the teacher ’ s output distribution to inherit nuanced decision boundaries and inter-class relationships. This “dark knowledge” enables small models to approach teacher-level performance while reducing memory and computational costs, which is critical for real-time applications especially when resources are insufficient. In natural language processing (NLP), KD has been widely adopted to deploy efficient systems without quality sacrifice, including in Machine Translation Quality Estimation (MTQE). Recent work extends KD beyond output imitation(Chi et al., 2025); (Labadie-Tamayo et al., 2025). While transferring reasoning abilities and intermediate rationales from LLMs to smaller models, they offer richer supervision for complex decision-making tasks. Aiming to enhance the performance of a model, knowledge distillation has common points with data augmentation. However, they operate on different principles. Data augmentation enriches the training set by generating new or modified samples, investing in greater input diversity. Knowledge distillation does not change the dataset itself; instead, it employs a teacher model ’ s outputs or intermediate representations to transfer task-specific knowledge in order to benefit the student model even from the same data.

1.2 Machine Translation Quality Estimation

Machine Translation Quality Estimation (MTQE) is the task of predicting translation quality without human reference translations. By providing segment-level quality scores, MTQE systems reveal translation quality for downstream tasks such as post-editing, informing users about reliability. State-of-the-art MTQE approaches, such as COMETKiwi (Rei et al., 2022c), apply large pretrained multilingual models to achieve strong correlation with human judgments. However, such models are computationally expensive and can be hard to deploy locally, making them unsuitable for some real-world scenarios. This motivates the exploration of KD to distill these capabilities into smaller, more efficient models, while maintaining high performance.

1.3 LLM as Annotator

Large Language Models (LLMs) have demonstrated notable ability as "LLM-as-a-judge" for translation evaluation (Kocmi and Federmann, 2023); (Fernandes et al., 2023). Through in-context learning, they can detect errors, explain their severity, and mimic human annotation quality. This enables the generation of high-quality synthetic supervision at scale, including fine-grained rationales together with frameworks like MQM. Proprietary LLMs such as GPT-4 have both upsides and downsides. In one hand, they offer state-of-the-art evaluation performance, in the other hand, their closed-source nature and high inference costs raise reproducibility and deployment concerns. Nevertheless, the LLMs as powerful annotators present a promising teacher in KD settings, capable of providing interpretive reasonings to the target scores.

1.4 Aim and relevance

The aim of this work is to develop and evaluate a reasoning-enhanced distilled model for MTQE, using an open-source small language model as the student and an LLM-generated rationale as auxiliary supervision. At the same time, the model supervised by human annotations is also trained and compared with the enhanced model, to validate the impact of the LLM reasoning in MTQE. The approach seeks to combine human-annotated gold scores with LLM-provided explanations, enabling the student model to learn not only the final quality assessment but also the scoring mechanism process behind it. This research is relevant to both academic and industrial contexts. For academia, it provides evidence on the role of explanatory signals in KD for regression-based NLP tasks, offering insights into when and how reasoning supervision improves model performance. For industry, it addresses the urgent need for efficient, high-quality QE models that can be deployed in translation pipelines, including scenarios with resource constraints or multilingual coverage requirements. Furthermore, it investigates the potential of LLMs to serve as scalable, high-quality annotators, contributing to the ongoing discussion on whether such models can complement or even compete with human evaluation in certain settings.

1.5 Problem Definition

1.5.1 Sub-problem 1: Why use LLM

LLMs excel in generating high-quality natural language explanations without retraining or special language settings, even for low-resource language or unseen domains. This flexibility allows the creation of rich synthetic training data without collecting costly human annotations for every scenario.

1.5.2 Sub-problem 2: Necessity of Knowledge Distillation

Although MTQE models have advanced significantly, their best-performing versions are still large and resource-intensive. Deploying such models in production pipelines faces constraints on inference speed, memory usage, and hardware requirements. KD offers a viable pathway to address these constraints, enabling the creation of smaller models that retain competitive performance while being faster, cheaper, and more accessible.

1.5.3 Sub-problem 3: Can LLM Compete with Human Annotators

Recent studies have shown that LLM-based evaluations demonstrate powerful abilities in MT assessment, particularly when guided by structured prompting (Kocmi and Federmann, 2023), (Kocmi et al., 2024). However, the reliability of LLM annotations depends on prompt design, model calibration, and bias control. When facing large scale dataset and complicated texts, would LLM always deliver stable high-quality assessments? With the help of augmented reasoning data, would the MTQE framework be enabled with significant performance improvement? This question will be discussed through a series of experiments.

1.6 Contribution

This thesis investigates whether a small, general-purpose open-source language model (Qwen3-0.6B) can be distilled into an effective QE model using LLM-generated reasoning as auxiliary supervision. The proposed approach uses human-annotated Z_{mean} scores and GPT-4o-mini–generated rationales as two complementary teachers. The student model is fine-tuned under two configurations, one using gold scores only and another concatenating gold scores with reasoning text. WMT 2021–2023 Direct Assessment datasets are used as data. Performance is compared across quality ranges, sentence lengths, and language pairs, with both correlation and error-based metrics. The main contributions are as follows：

- A reasoning-enhanced distillation strategy that combines GPT-generated reasoning texts with gold human scores to guide student model training.

- A systematic comparison between gold-score-only supervision and reasoning-augmented supervision, evaluating their impacts across correlation and error-based metrics.

- An error analysis revealing the complementary strengths of each supervision type, particularly in extreme quality ranges and across sentence lengths.

1.7 Outline

Chapter 2 reviews related work in MTQE, KD, and LLM-based evaluation. Chapter 3 details the datasets, preprocessing steps, and how auxiliary reasoning data is generated. It also presents the methodology, including the distillation strategy, model architecture, and training configuration. Chapter 5 reports all the experimental results including fresh out-of-box Qwen models, the gold-score-supervised model, and the distilled model. Their performances are compared with the state-of-the-art COMETkiwi to situate this study in a professional domain. Chapter 6 provides quantitative and qualitative error analyses based on the predictions of the two main models. The impact of auxiliary GPT reasoning supervision is carefully examined. Chapter 7 discusses implications and limitations, with directions for future research. Chapter 8 concludes this study.

Chapter 2

Related Work

Machine Translation Quality Estimation (MTQE) is the task of automatically assessing the quality of a machine-generated translation without access to a human-authored reference translation. This reference-free nature makes QE an indispensable tool in real-world translation workflows, where it can be used to inform end-users about the reliability of a translation, determine whether a segment requires human post-editing, or filter out low-quality content before it reaches the user. The field has evolved significantly, moving from early statistical methods that relied on handcrafted features to the current paradigm dominated by large-scale neural networks fine-tuned on human judgments.

Knowledge Distillation (KD) has been implemented as the main technique to address this challenge. First formalized by Hinton et al. (2015), KD is a model compression method where a compact ”student” model is trained to mimic the behavior of a large, high-performing ”teacher” model. This process transfers the rich ”knowledge” learned by the teacher, enabling the student to achieve comparable performance with less computational resources. This literature review provides a systematic analysis of the key research works supporting the application of KD for MTQE.

2.1 From Feature Engineering to Neural Frameworks

The initial phase of QE research was characterized by systems that were built on handcrafted features. These approaches involved extracting a wide array of linguistic indicators, such as source and target sentence lengths, language model perplexity, and word alignment statistics. By feeding them into traditional machine learning models like Support Vector Machines (SVM), a quality score was predicted. These kind of models were limited by the range of covered features, resulting in models that failed to adapt themselves to specific scenarios where features are hard to extract or too complex to represent certain characteristics of the data.

The rise of deep learning, especially large pre-trained language models (PLMs), brought big changes to the field. To support research in this new approach, Kepler et al. (2019) introduced OpenKiwi, a comprehensive PyTorch-based open-source framework. OpenKiwi standardized experiments by including the winning systems from the WMT 2015–18 quality estimation contests, supporting both word-level and sentence-level tasks. The key architecture was the Predictor-Estimator model, which breaks down the QE task into token-level error identification (the predictor) and sentence-level score aggregation (the estimator), offering a strong guidance to the learning process.

Building upon the success of PLMs, Ranasinghe et al. (2020) developped TransQuest, a simple yet powerful QE framework based on cross-lingual transformers. TransQuest demonstrated that by leveraging strong pre-trained embeddings from models like XLM-R(Conneau et al., 2020), it was possible to reach state-of-the-art performance without relying on complex, hand-crafted model architectures. Outperforming the OpenKiwi baseline, this framework investigated two setups: a single-encoder model (MonoTransQuest) that processes the source and translation as a single sequence, and a siamese model (SiameseTransQuest) that encodes them separately and compares their embeddings.

The state-of-the-art model in MT evaluation has been largely defined by the COMET framework, introduced by Rei et al. (2020). COMET builds on recent progress in cross-lingual pretrained language models. It produces multilingual evaluation systems that align closely with human judgments. These models are adaptable and effective across language pairs. A shared encoder (e.g., XLM-R) is used to embed the source, machine translation, and reference into a common space, from which a quality score is predicted. Although reference-based, the framework is highly flexible and could be adapted for reference-free QE tasks. Models such as COMET-22 and COMETKiwi are specifically trained for the QE task by omitting the reference translation directly from the input, relying on the source-translation pair to predict quality(Rei et al., 2022a), (Rei et al., 2022c). Being highly effective, these QE models are used as reward signals in reinforcement learning pipelines to improve MT systems.

Given the increasing power of reference-free models, researchers started to think about the necessity of human references. Rei et al. (2021) asked: "Are References Really Needed?" in their submission to the WMT 2021 Metrics Shared Task. The paper has provided compelling evidence that reference-free COMET models were becoming highly competitive with the reference-based ones, outperforming the best reference-based COMET model from the previous year (on development data). This finding suggests that for many practical applications, the expensive process of creating human reference translations could be replaced by powerful QE models. The same work also introduced COMETinho(Rei et al., 2022b), a lightweight distilled version of COMET that is 19 times faster while maintaining SOTA correlation with human judgments, highlighting the importance of efficiency.

2.2 Human Evaluation: The Ground Truth for Supervision

Supervised QE models rely heavily on human-labeled data. Human evaluation appears as the principle standard, and the way these judgments are collected continues to evolve. Direct Assessment (DA) is a widely used method, where annotators assign a single overall quality score on a scale (e.g., 0–100). DA is quick and low-cost, but it often lacks detail. The scores are subjective and do not reveal specific translation errors.

To address this, the Multidimensional Quality Metrics (MQM) framework has become the main approach for detailed human evaluation. It asks expert annotators to mark error spans, label them with error types (like accuracy or fluency), and judge how serious each error is. This creates structured data that helps train more precise QE models. But compared to Direct Assessment (DA), MQM is slower and more costly, as it depends on highly trained experts.

To balance MQM’ s detail and DA’ s scalability, Kocmi et al. (2024) introduced Error Span Annotation (ESA), a hybrid method. In ESA, annotators first highlight error spans and rate their severity as minor or major. Then, they give an overall DA score for the segment. This two-step process helps to give the overall score considering concrete errors. Validation results showed that ESA is faster and cheaper than full MQM, while still ranking MT systems with similar accuracy and consistency. It offers a practical path toward scalable, high-quality human-labeled data.

2.3 Explainable and Fine-Grained Evaluation

The limited aspect of traditional neural QE models is that they only give a single score, without information about the errors in the translation. This lack of clarity has led to more focus on explainable evaluation. xCOMET has been used to address this problem. It extends the COMET framework by predicting both a sentence-level score and word-level error spans with severity tags (OK, Minor, Major, Critical). This multi-task setup gives detailed feedback, helping users understand both the quality and the problem of a translation, making the evaluation more transparent(Guerreiro et al., 2024).

The appearance of Large Language Models (LLMs) has introduced new, powerful paradigm for MT evaluation, often termed ”LLM-as-a-Judge”. Instead of fine-tuning a model on QE data, this method relies on the vast world-knowledge and reasoning capabilities of LLMs through in-context learning and prompting. In the work of Fernandes et al. (2023), the researchers proposed AutoMQM, which guides the LLM to detect and classify translation errors in the MQM style instead of giving a score directly. The labeled error spans are then used to compute a final quality score. This technique improves performance of the language models, and enables the output to be fully interpreted as well by producing error annotations that are very close to human judgments.

Similarly, Kocmi and Federmann (2023) introduced GEMBA-MQM, a metric that uses GPT-4 with a three-shot prompting technique to detect translation quality error spans in a reference-free setting. The use of language-agnostic prompts is a key feature. Early results demonstrated that this method achieves state-of-the-art accuracy in ranking MT systems, validating the potential of LLMs for fine-grained QE tasks. However, the fact that it depends on a proprietary black-box model limits reproducibility and transparency.

2.4 Alternative Formulations

Most QE research has focused on a regression framework to predict an absolute score. Moosa et al. (2024) has proposed MT-Ranker, which converts reference-free MT evaluation as a pairwise ranking problem. The model’s task is to predict the better one between competing translations to the same source sentence. This ranking-based approach more closely imitates how humans intuitively compare translations. It achieves stronger alignment with human judgments, without relying on direct human-labeled training data. Instead, it learns from weak supervision using synthetic examples and draws indirect signals from related tasks, such as natural language inference.

Finally, another important dimension of QE is acknowledging and quantifying the model’s own uncertainty. Neural metrics are trained on human scores, which can be noisy and inconsistent. But these models usually give just one fixed number as output, which hides the extent of confidence the model has about the quality. To solve this, Glushkova et al. (2021) has introduced an uncertainty-aware evaluation framework, which is an extension of COMET with techniques such as Monte Carlo dropout and deep ensembles. Rather than returning just a single score, the model outputs a score along with a confidence interval, which shows how confident the model is for that prediction.

2.5 Knowledge Distillation Motivation and Mechanism

Knowledge Distillation (KD) has become a widely used technique for model compression, aiming to transfer the capabilities of a large, high-performing ”teacher” model to a smaller, more efficient ”student” model (Gajbhiye et al., 2021), (Mirzadeh et al., 2020). The paradigm, introduced by Hinton et al. (2015), involves training the student to mimic the teacher’s output distribution (also called the soft labels) by minimizing the Kullback-Leibler (KL) divergence between them (Huang et al., 2022). This process allows the student to learn both the correct predictions and the nuanced inter-class relationships captured by the teacher, which are considered a valuable form of ”dark knowledge” (Tang et al., 2020). There is a practical reason for this: state-of-the-art models, particularly in natural language processing, are often too large and computationally expensive for real-time deployment in resource-constrained environments in real-world. This is especially true for Machine Translation Quality Estimation (MTQE), as its recent success relies on large multilingual pre-trained models. This fact makes knowledge distillation a critical tool for creating lightweight, yet competitive QE systems (Gajbhiye et al., 2021).

To better understand the mechanism of knowledge distillation, Tang et al. (2020) deconstructed the teacher’s knowledge into three hierarchical levels. They argue that KD’s effectiveness comes from: (1) a regularization effect similar to label smoothing, which they described as knowledge of the ’universe’; (2) domain knowledge that comes from the teacher’s learned class relationships; and (3) the output layer geometry based on the teacher’s confidence in its predictions. This analytical framework helps to explain many distillation cases, both the success ones and failures.

2.6 Capability Gap in Distillation

An important and counter-intuitive challenge in KD is the observation that a stronger or significantly larger teacher does not always lead to a better student. In contrast, it lowers the student’s performance sometimes, compared to training with a weaker teacher (Huang et al., 2022), (Mirzadeh et al., 2020). This phenomenon is often described as a ”capacity gap,” where a small student model lacks the capacity to perfectly replicate the complex output distribution of a much larger and more powerful teacher. When the student cannot bridge this gap, the strict object to minimize the KL divergence could be an obstacle, leading to suboptimal training.

To address this, Mirzadeh et al. (2020) proposed a multi-step distillation framework that introduces an intermediate-sized model, or ”Teacher Assistant” (TA). The TA first learns from the large teacher and then, in turn, teaches the final smaller student. This approach bridges the capacity gap effectively by breaking the whole knowledge transfer process into smaller and manageable steps, demonstrating that a teacher can transfer its knowledge to a smaller student more effectively.

The stronger teacher problem is addressed elsely by a different line of research, by modifying the distillation objective itself to avoid the strict replication of the teacher's output distribution. Huang et al. (2022) argue that when the discrepancy between teacher and student predictions is severe, forcing an exact match via KL divergence can disturb the student's training. They argue that for knowledge transfer, it is sufficient and effective to preserve the teacher's preference or ranking of classes and their relative relationships within the predictions. To address this issue, they introduced DIST, using a loss based on correlation to capture both the inter-class relations in each example and the intra-class relations across different examples. This avoids the tough need to match absolute values for the student, and instead guides it to learn relational knowledge revealed in the teacher’ s predictions. This idea of highlighting the relative judgments could be seen in other domains, such as the work by Moosa et al. (2024) on MT-Ranker, which frames machine translation evaluation as a pairwise ranking task. According to their idea, learning from relative comparisons could yield more stable outcome and is easier for the model to learn than learning from absolute scores.

2.7 Distilling Reasoning Abilities and Supervision with Rationales

As Large Language Models (LLMs) are rising, the focus of distillation has shifted from mimicking predictions to transferring complex, multi-step reasonings. This shift has been motivated by the observation that abilities like chain-of-thought (CoT) reasoning typically tend to appear in large-scale models with hundreds of billions of parameters. Fu et al. (2023) introduced the concept of model specialization, showing that such complex reasoning abilities can be distilled from large models such as GPT-3.5 into much smaller ones like T5 variants, with fewer than 11 billion parameters. They assume that by concentrating a small model's focus on a specific target task, such as multi-step math problems, it could make more efficient use of its limited capacity, resulting in strong task-specific performance.

One way to apply this distillation of reasoning is to use the teacher LLM to generate intermediate reasoning steps or rationales, which then serve as additional supervision for the student beside the core predictions. Hsieh et al. (2023) proposed distilling step-by-step, a simple methodology where a teacher's rationales are used to fine-tune a smaller student model. This approach was shown to be highly data-efficient, enabling a 770M parameter model to outperform a 540B parameter LLM using only a small portion of the available training data. Expanding this idea, Samarinas and Zamani (2025) combined knowledge distillation with reinforcement learning for the task of document re-ranking. In their research, a teacher LLM first generates a dataset of query-document pairs along with relevance explanations. A smaller student model is then trained to replicate the reasoning and labeling capability through supervised fine-tuning, after which reinforcement learning is used to further refine the student's ability to generate high-quality explanations. Together, these works demonstrate a powerful new aspect of KD, which focuses on transferring the cognitive processes of LLMs, but not just the final predictions.

Larionov et al. (2024) proposed xCOMET-lite, a small and efficient metric for MT evaluation. Their methodology involves training a lightweight DeBERTa-based student model(He et al., 2021) to replicate the outputs of the large xCOMET-XXL teacher, using a large-scale synthetic dataset of 14 million translation-reference pairs. Additionally, they explore quantization (GPTQ, LLM.int8(), QLoRA) and structured layer pruning with parameter-efficient fine-tuning, finding that quantization effectively reduces model size with minimal performance degradation. The key contributions of the work include the development of a black-box distillation pipeline for learned quality estimation, validating that xCOMET-lite retains over 92% of the teacher's performance with only 2.6% of its parameters. Their work has shown that distillation combines well with quantization but not with pruning. This work highlights a viable path toward deploying high-quality evaluation metrics in resource-constrained settings.

Low-Rank Adaptation (LoRA) (Hu et al., 2021) is a fine-tuning approach that updates only small low-rank matrices inserted into a model' s layers, leaving the original weights unchanged. This design greatly reduces the number of trainable parameters, lowering both memory usage and training cost, while preserving model quality. LoRA can be combined with other efficiency methods such as quantization and supports maintaining multiple lightweight adapters for different tasks. In this study, LoRA enables efficient fine-tuning of the Qwen3-0.6B student model for MTQE under limited computational resources.

2.8 Qwen Family and Qwen3

The Qwen family has evolved from early open-weight LLMs into a broad suite spanning text-only, vision-language, audio, long-context, coding, multimodal "omni," and latest-generation models. The first Qwen release established strong base and instruct LLMs, quickly extended to multimodal variants such as Qwen-VL and Qwen-Audio.(Bai et al., 2023a); (Chu et al., 2023)

In 2024, Qwen2 expanded sizes (sub-billion to tens of billions) and introduced MoE options and multilingual gains, while specialty lines (Qwen2-VL and Qwen2-Audio) upgraded perception and audio interaction.(Bai et al., 2023b); (Chu et al., 2024) Qwen2.5 pushed scale and post-training, adding 1M-token context variants, coder models, and an "omni" stack for text-image-audio-video, alongside a stronger VL series((Qwen et al., 2025)).

In 2025, Qwen3 unified these advances with a refreshed lineup from lightweight (0.6B) to massive models, designed to cover a wide range of sizes from lightweight variants to multi-billion-parameter models (Yang et al., 2025). It builds on advancements in multilingual training, instruction tuning, and alignment to deliver strong performance across reasoning, code generation, and multilingual understanding tasks. Qwen3 models adopt an optimized Transformer architecture with efficiency-focused improvements, enabling competitive performance while maintaining scalability for both research and deployment. In this work, the 0.6B parameter version, Qwen3-0.6B, is selected as the student model due to its balance between capability and computational efficiency, making it suitable for knowledge distillation in MTQE under constrained resources.

2.9 Method Overview

While existing work has explored a variety of frameworks for MT quality estimation and different strategies for knowledge distillation, many approaches rely on large teacher models, complex training pipelines, or task-specific architectures. In contrast, this work investigates whether a small, general-purpose open-source language model can be distilled into an effective QE model using simple supervision signals. Specifically, two types of supervision are considered as signals: (1) human-provided quality scores, and (2) GPT-generated reasoning about translation quality, which offers a form of indirect guidance. This setting allows to experiment with reasoning-based distillation without depending on large proprietary teachers. In the next sections, I present the dataset used for training and evaluation, followed by the details of the distillation experiments using Qwen3-0.6B as the student model.

Chapter 3

Methodology

3.1 Background

This study investigates the use of knowledge distillation (KD) to improve the machine translation quality estimation (MTQE) ability of a small open-source language model, Qwen3-0.6B (referred as “the student model”)(Yang et al., 2025). Although pretrained on large-scale general corpora, the student model hasn’ t been trained explicitly for MTQE tasks. My objective is to enhance its scoring capability on translation pairs without relying on reference translations through a series of knowledge distillation experiments.

Developed by Alibaba Cloud, Qwen3-0.6B represents the smallest variant in the latest generation of the Qwen large language model series. In the knowledge distillation-inspired setup, Qwen3-0.6B acts as the student, tasked with learning the complex nuances of translation quality assessment from the provided gold-standard scores and auxiliary data. The selection of this model is guided by three reasons: efficiency, performance, and accessibility. As a Small Language Model (SLM), it offers a lightweight and computationally efficient alternative to larger models, making it practical for local deployment and real-world applications. This aligns with the core principle of this study, aiming to compress knowledge into a more accessible form. Despite its compact size, the Qwen series has demonstrated robust performance across a wide range of downstream tasks, showing strong multilingual abilities, making it a suitable candidate for learning the subtle patterns required for quality estimation. The model is released under the Apache 2.0 license which ensures full access to model weights and architecture, essential for transparent and reproducible academic research.

The Qwen3-0.6B model is built upon a decoder-only Transformer architecture comprising 28 layers and approximately 0.6 billion parameters (0.44B non-embedding). It employs Grouped-Query Attention (GQA) with 16 query heads and 8 key-value heads to enhance inference efficiency. A key feature for this task is its extensive 32,768-token context window, which is crucial for processing the concatenated source, translation, and supplementary reasoning texts without truncation. The model’s strong foundation is derived from an extensive pretraining phase on nearly 36 trillion tokens spanning 119 languages. This diverse dataset is specifically curated to enhance the model’s reasoning and multilingual capabilities, making it an ideal base for translation quality estimation tasks.

For the stronger teacher model, I chose ChatGPT, a large proprietary language model, as the teacher model. The teacher model exhibits strong reasoning and explanatory power in many tasks, and can be used as a good annotator of translation quality(Kocmi and Federmann, 2023). The core idea is to leverage GPT's interpretations to guide the student model towards more accurate quality judgments. I selected GPT-4o-mini for its combination of efficiency and powerful multilingual reasoning, which enables it to analyze diverse language pairs without special settings and produce coherent explanations. I prepare a dataset of source (S_{src}) and machine-translated (S_{mt}) sentence pairs annotated by human experts with quality score. Each pair together with its score are input to GPT, to generate natural language reasoning explanations about why this translation has gained such a score. Then, I concatenated the (S_{src} + S_{mt}) pair with the GPT generated reasoning texts as an input into the student model. The training objective is to minimize the prediction loss against the human-annotated quality score with the help of GPT-generated reasoning. The GPT-generated reasoning texts surve as an augmented data. By injecting this data into the student model, I expect that GPT would present a good reason to why the sentence pairs have obtained such a quality score from the human annotators, and pass this scoring mechanism to the student as a form of knowledge distillation.

In order to compress the model to facilitate the deployment in real world, I used LoRA(Hu et al., 2021). This technique freezes the majority of the model's original parameters, and injects small trainable "adapter" matrics into specific layers of the architecture. In the training process with Unsloth, a quantized 4-bit version of Qwen3-0.6B is used to reduce the precision of the weights from the standard 16 or 32 bits down to 4 bits, drastically cutting the memory usage and speeding up computation.

The aim of this approach is to combine the student model's general language competence with the task-specific interpretability from the teacher to get a stronger model that is suitable for MTQE task. The hypothesis is that GPT’ s reasoning supervision will provide additional semantic guidance to the student model, allowing it to internalize patterns relevant to translation quality assessment, and eventual, obtain the scoring ability that is close to human annotation.

It is essential to compare the results against a recognized state-of-the-art model to properly measure the performance of the trained models. For this role, I selected a model from the COMET (Cross-lingual Optimized Metric for Evaluation of Translation) framework. COMET is a learned model trained specifically on translation QE tasks, making it the de facto standard for MTQE research. It is a necessary step to compare the trained model with COMET to validate whether the training is effective and competitive. I selected wmt22-cometkiwi-da, a powerful reference-free QE system developed by Unbabel. The performance of the baseline is presented together with other trained models in Chapter 4.

3.2 Data

To investigate the effectiveness of knowledge distillation in Machine Translation Quality Estimation (MTQE), I utilize the publicly available human Direct Assessment (DA) data from the WMT 2021–2023 shared tasks on Quality Estimation. These datasets are widely used benchmarks in QE research, providing sentence-level human scores for a diverse set of language pairs.

3.2.1 WMT Datasets

The Workshop on Machine Translation (WMT) Shared Task is an annual competition that plays a central role in the machine translation (MT) research community. Each year, WMT releases benchmark datasets primarily consisting of parallel corpora from the news domain, supplemented by other sources. These datasets support both the training and standardized evaluation of MT systems across a wide range of language pairs. The consistent release over the years remains a key resource for the development and assessment of translation models, enabling comparisons across different MT systems.

In this work, I focus on the Direct Assessment (DA) datasets from WMT 2021, 2022 and 2023. The decision is motivated by three main factors. First, these are the most recent WMT datasets, ensuring compatibility in annotation style. Second, all the years offer high-quality DA annotations across a diverse set of language pairs, including several low-resource directions, which align with the objective of training robust multilingual QE models in this research. Third, earlier datasets (e.g., WMT 2017–2019) follow a different annotation scheme (HTER scores), which focus on post-editing estimation rather than the quality of the translation. This limits the usage of the earlier datasets in this research. WMT21 adopts largely the same datasets as WMT20, with the addition of one extra language pair, making the inclusion of WMT20 redundant for the experiments. In the following table, the language pairs and sample sizes are clearly presented.

Year	**Language Pairs**	**Total Sentences**
2021	en-de, en-zh, ne-en, ro-en, ru-en, si-en	42,000
2022	en-de, en-mr, en-zh, et-en, ne-en, ro-en, ru-en, si-en	74,762
2023	en-gu, en-hi, en-ta, en-te	28,000
Total	17 Unique Language Pairs	144,762

Table 3.1: Statistics of dataset by year and language pairs

3.2.2 Data Structure and Target Variable

Each example in the WMT datasets is a sentence-level annotation containing several fields. For this research, three key columns are used: src (the original source sentence), mt (the machine-translated hypothesis), and Z_{mean}. The Z_{mean} score is the mean of standardized scores (z_scores) from three human annotators. This normalization process mitigates inter-annotator variability and biases, resulting in a more reliable and robust quality. The Z_{mean} score is adopted as the ground-truth "gold" label in trainings. The Z_{mean} scores in this aggregated dataset range from -7.54 to 8.45, with a mean of -0.0016 and a standard deviation of 0.8473, approximating a normal distribution.

3.2.3 Data Preprocessing

Although the WMT datasets are generally well-structured, I identified minor instances of data corruption within the files, such as entire rows concatenated into a single column and contents from different columns squeezing together. To ensure data integrity, I developed and executed an automated parsing script to detect and isolate these malformed rows based on abnormal token counts (e.g., excess tab or newline characters).

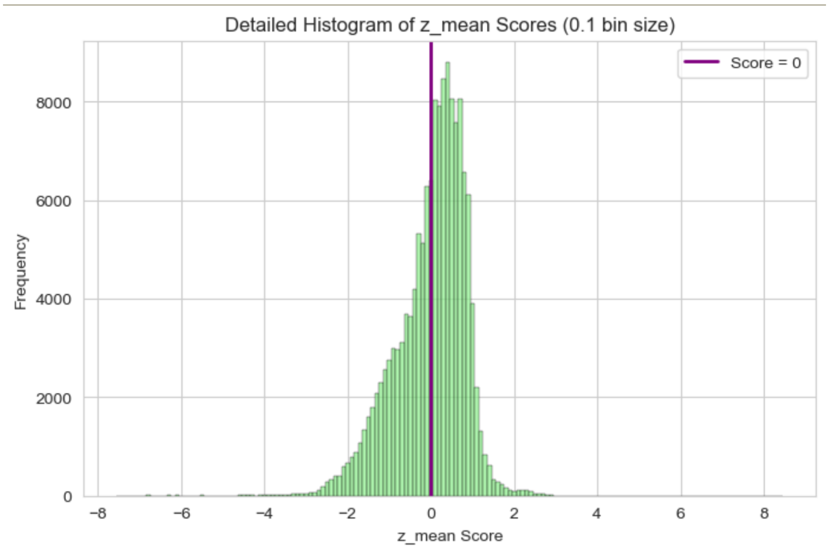

Figure 3.1: Range of the gold scores

These isolated rows were then manually corrected and reinserted into the dataset. I retained all naturally occurring linguistic noise, such as punctuation errors or grammatical imperfections, as these artifacts are authentic features of machine translation output and are crucial for training the model to distinguish between high and low quality translations.

3.2.4 Data augmentation: GPT-generated reasoning

In addition to the Z_{mean} score, I introduce a second, richer form of supervision derived from a large language model. For each S_{src}, S_{mt} pair in the dataset, I prompted GPT-4o-mini to generate a textual explanation justifying the given Z_{mean} score. This process yields natural language reasoning texts that identifies specific errors or strengths in the translation, providing context for the quality score to each sentence pair. This generated text serves as ”soft explanatory knowledge” and is a core component of the reasoning-enhanced distillation strategy, allowing the student model to learn not only **what** the quality score is, but also **why**. The following is an example of GPT-generated reasoning to a $S_{src} + S_{mt}$ pair:

- *Source*: Despite this Mithridates joined Antiochus Hierax against Seleucus.
- *Mt*: 尽管如此，米特里达斯还是加入了反对塞卢库斯的安蒂奥丘斯 · 希拉克斯。
- *GPT Reasoning*: This translation merits a score of 2 because it faithfully renders “Despite this Mithridates joined Antiochus Hierax against Seleucus”into idiomatic Chinese with correct transliterations of 米特里达斯, 安蒂奥丘斯 · 希拉克斯 and 塞卢库斯, preserves the causal concession and action relationship, and reads fluently; the only minor blemish is unconventional spacing around punctuation, which does not impede understanding.

As shown, the reasoning text is a short paragraph in natural language around 90-120 words. It is stored with each $S_{src} + S_{mt}$ pair in a same TSV file.

3.3 Task Formulation

The primary goal of this study is to develop an efficient machine translation (MT) quality estimator. This task is framed as a supervised, generative regression problem. The goal is to fine-tune a small open-source language model, specifically ”Qwen3-0.6B”, to predict quality scores for a dataset which consists of a large amount of sentences and its corresponding machine translation. The scores should closely align with human-annotated gold standard scores, enabling the model to serve as a reliable, reference-free quality evaluation tool.

To investigate the role of knowledge distillation in this process, I formulate two distinct training setups. The input for both models consists of a source sentence (S_{src}) and a machine-translated sentence (S_{mt}). The training objective is to predict the gold standard quality score, denoted as Z_{mean}. For the second model, auxiliary explanatory data is used to assist the model’ s understanding of the translation quality.

- **Student model fine-tuned with $Z_{\textbf{mean}}$ (Baseline)** is trained using only the core data: the (S_{src}, S_{mt}) pair and its corresponding Z_{mean} score.
- **Student model fine-tuned with GPT reasoning (Enhanced)** is trained on an augmented dataset that includes the (S_{src} + S_{mt} + S_{GPT}) triplet. This reasoning provides a clear rationale for why a given translation received its specific quality score.

In both configurations, the model is expected to generate a single numerical value as its output, representing the predicted quality score. For the second model, the GPT reasoning serves as supplementary context to guide the learning process, rather than as a pseudo-label to be replicated. The learning objective is to minimize the gap between the model’s predicted score (y_{pred}) and the true gold standard score (Z_{mean}).

During training, the inputs are structured into a specific prompt format. This format concatenates the source sentence, the machine translation, and, for the enhanced model, the auxiliary reasoning. The model is then instructed to process this contextual input and generate the final numerical quality score. This prompt-based approach enables a language model to convert the generation to a regression task.

To comprehensively evaluate the performance of the student models, I utilize a set of four standard metrics. The Pearson correlation coefficient (r) is considered the primary metric(Akoglu, 2018). It measures the linear relationship between the predicted scores and the gold Z_{mean} scores on the test set, indicating how well the model’s predictions align with human judgments. Ranging from -1 to 1, its different value range signifies different quality buckets:

- Values close to ±1 indicate a perfect linear relationship. If above ±0.7, the correlations are consideres strong.
- Values between ±0.3 and ±0.6 indicate weak to moderate associations.
- When $|r|$ is below 0.1, the correlation is negligible, while 0 indicates zero correlation.

A higher Pearson r signifies a more reliable model as estimator. Mean Absolute Error (MAE) is also reported, which provides a direct and interpretable measure of the average prediction error. Mean Squared Error (MSE) and Root Mean Squared Error

(RMSE) are also used. These metrics are particularly sensitive to large deviations, with RMSE being in the same unit as the original scores, offering a clear view of the magnitude of significant prediction errors.

3.4 Student model

The core methodological step in this study is adapting Qwen3-0.6B, a generative language model to perform a numerical regression task. I achieved this by framing the task as instruction-following through a prompt-based fine-tuning process. The model is presented with a structured input, including a system message defining its role and a user message containing the source and translated sentences. It is then instructed to generate only a numerical quality score. Through supervised fine-tuning, the model learns to map the semantic and syntactic features of the input text to a single floating-point number. The student model architecture is then fine-tuned under two distinct conditions, one with and one without auxiliary reasoning texts, to directly evaluate the research hypotheses.

3.5 Distillation strategy

3.5.1 Conceptual Framework

To train the student models, I employ a framework inspired by knowledge distillation. The actual approach is different from classic distillation methods, which typically involve matching the logits (soft labels) of a large teacher model. Instead, I conceptualize distillation as the transfer of structured knowledge from two complementary sources: human-annotated scores and GPT-generated reasoning. The primary goal is not to teach the student model to replicate the reasoning text itself, but to use this text as auxiliary supervision. This supplementary information is designed to guide the student model to look into the specific errors in the translations more effectively.

3.5.2 Defining the ”Teacher” and the ”Knowledge”

The ”teacher” in the framework is a composite entity. The first component consists of human experts, who provide the ground-truth knowledge in the form of Z_{mean} scores. This represents the definitive ”what”: the final, authoritative judgment on translation quality. The second component is a ”reasoning teacher”, for which I utilize GPT-4o-mini. This model acts as an explicator, generating the ”why” behind a score by providing a detailed rationale. This process facilitates the transfer of two distinct types of knowledge:

- **Hard Target Knowledge**: The numerical Z_{mean} score, which serves as the direct and unambiguous regression target, directly reflecting the translation's quality.

- **Soft Explanatory Knowledge**: The GPT-generated reasoning texts, which are treated as supervisory information during training. It enriches the input data, enabling the student model to more precisely grasp the underlying mechanics of the scoring process.

3.5.3 The Mechanism of Knowledge Transfer

The mechanism for knowledge transfer is implemented through a comparative experimental design which allows to isolate and measure the impact of the explanatory knowledge. I trained the Qwen3-0.6B student model under two distinct conditions:

- **Condition 1: Gold-score-supervised training.** In this setup, the student model is trained using only the S_{src} + S_{mt} pairs and their corresponding Z_{mean} scores. This represent a standard supervised learning approach where the model must infer the principles of quality estimation entirely on its own from the numerical targets.

- **Condition 2: Reasoning-Enhanced training.** Here, the student model is trained on an augmented input format that includes the S_{src} + S_{mt} + S_{GPT} triplet, while still being tasked to predict a quality score for each triplet. The reasoning text is concatenated directly into the input prompt, providing detailed context for why the target score is given. The results from this enhanced model are expected to demonstrate an improved understanding of the scoring mechanism, guided by the insights from the reasoning teacher.

3.5.4 The Role of Reasoning as a Learning Signal

In this framework, the reasoning text is conceptualized not as mere supplementary data, but as a structured learning signal. By processing the rationale alongside the source and translated sentences, the student model is encouraged to learn the causal link between specific linguistic phenomena (e.g. lexical choice errors, grammatical inaccuracies, and stylistic inconsistencies, etc.) and the final numerical quality score. I hypothesize that this approach could force the model to move beyond surface-level pattern matching and gain deeper insight into the scoring mechanism. Finally, the model should be able to learn effectively **why** a translation is good or bad.

The reasoning texts from GPT are generated using a combination of a system prompt and a user prompt. The system prompt was designed to be highly explicit, first assigning the model the persona of a ”multilingual translation evaluation expert”. It then defined the task clearly: to predict a numerical quality score for a given source-target translation pair within a specified range of -2.0 to 2.0. To guide the model, it also included descriptive scoring criteria for different score intervals (e.g., -2.0 to -1.0 for ”very poor translation”). The prompt ends with a direct instruction to ”output only the numerical score” to constrain the output format.

```
You are a multilingual translation evaluation expert. Your task is to
predict the quality score for translation pairs. The quality score
is a number between -2.0 and 2.0 and could fall into the following
categories: -2.0 to -1.0: Poor or very poor translation with meaning
deviation or severe errors. -1.0 to 0.0: Flawed translation but
understandable. 0.0 to 1.0: Good translation in general. 1.0 to 2.0:
Excellent or flawless translation. The relative position of the scores
in this range indicates subtle differences in translation quality.
Based on the provided source, translation, and reasoning, output only
the numerical score.
```

The user prompt was a simple and clear concatenation of the source and machine-translated sentences, formatted as:

```
Source sentence:\n{source_text}\n\n### MT sentence:\n{mt_text}
```

3.5.5 Learning Objective and Loss Function

The final learning objective and the mechanism for loss computation remain the same in both experimental conditions. The model's performance is optimized by minimizing the Mean Squared Error (MSE) between its generated numerical score and the Z_{mean} gold score. Therefore, the distillation strategy does not alter the fundamental loss function. The innovation lies in enriching the information available to the model during the training process. I expect that the enriched context will enable the reasoning-enhanced model to achieve a lower final loss and more importantly to demonstrate superior generalization capabilities when evaluated on test set.

In this work, the model is trained with cross-entropy loss, a widely used objective function for classification and language modeling tasks. It measures the divergence between the predicted probability distribution over tokens and the true distribution, penalizing the model more heavily when high-probability predictions diverge from the target. When adapting the model to predict a scalar quality score, the output token sequence (representing the score) is generated through the original generative head, and the corresponding loss is computed over its constituent subtokens using cross-entropy. This is different from a direct regression task, where the model is optimized with a regression loss such as mean squared error (MSE).

3.6 Experimental Setup

3.6.1 Training Framework and Optimization

To implement the designed fine-tuning process, I utilized the SFTTrainer class from the Hugging Face TRL (Transformer Reinforcement Learning) library. This trainer was selected as it provides a high-level, streamlined API specifically designed for the supervised fine-tuning of generative language models. A critical feature of the SFTTrainer[1] that was essential for this task is the *completion_only_loss* parameter. By enabling this feature, I instructed the trainer to compute the loss exclusively on the generated numerical score (the "completion") rather than on the entire input prompt. Since the prompts are significantly longer than the target output, calculating loss across the entire prompt would risk shifting the model's learning objective towards replicating the prompt structure itself, thereby fail the experiments.

To further enhance training efficiency, I integrated the Unsloth[2] library into the framework. Unsloth significantly accelerates the training process and reduces memory consumption by leveraging manually implemented Triton kernels for key operations within the Transformer architecture. This optimization typically results in a theoretical 2x speedup in training and can largely decrease GPU memory usage, which was vital in making the fine-tuning process feasible on resource-constrained hardware.

In line with the goal of efficiency, I employed a Parameter-Efficient Fine-Tuning (PEFT) strategy on the main models instead of full fine-tuning. I utilized a 4-bit quantized version of the base model, **Qwen3-0.6B-unsloth-bnb-4bit**. On top of this quantized base, I applied Low-Rank Adaptation (LoRA)(Hu et al., 2021). By only

[1] https://huggingface.co/docs/trl/sft_trainer

[2] https://unsloth.ai

training these low-rank adapters, I adapt the model to quality estimation task with a tiny fraction of the computational cost of full fine-tuning. For the LoRA configuration, I set the rank of the decomposition matrices to 32, which balances expressive power and the number of trainable parameters. The *lora_alpha* was also set to 32, acting as a scaling factor for the learned weights. I targeted all key linear layers within the Transformer blocks for adaptation, including the query, key, value, and output projections (*q_proj*, *k_proj*, *v_proj*, *o_proj*), as well as the feed-forward network layers (*gate_proj*, *up_proj*, *down_proj*), to ensure comprehensive task adaptation.

3.6.2 Hyperparameter Configuration

The hyperparameters for the training process were carefully selected to ensure stable convergence and effective learning. All configurations were kept consistent across both the baseline and reasoning-enhanced experimental conditions to allow for a fair comparison. The key hyperparameters used for all training runs are detailed in the following table. The effective batch size was 16, achieved by setting the per-device batch size to 4 and accumulating gradients over 4 steps. I utilized the *paged_adamw_32bit* optimizer for memory-efficient training and a learning rate of 2e-5 with a warmup phase of 100 steps to stabilize the initial stages of training.

Hyperparameter	**Value**
`per_device_train_batch_size`	4
`gradient_accumulation_steps`	4
`num_train_epochs`	3
`learning_rate`	2e-5
`bf16`	True
`optim`	paged_adamw_32bit
`warmup_steps`	100
`logging_steps`	50
`eval_strategy`	steps
`eval_steps`	1000
`save_strategy`	steps
`save_steps`	1000
`load_best_model_at_end`	True
`metric_for_best_model`	eval_loss
`save_total_limit`	1
`completion_only_loss`	True
`group_by_length`	True

Table 3.2: Hyperparameters used for model training.

3.6.3 Platform and Hardware

The experiments were conducted across two distinct cloud computing platforms. The initial phase of generating reasoning data with GPT-4o-mini was performed on Amazon Web Services (AWS), a leading cloud platform providing scalable computational resources. Specifically, I utilized a g5.xlarge instance, which is equipped with an NVIDIA A10G GPU. The primary model training and fine-tuning phase was conducted on AutoDL, a specialized GPU rental platform tailored for deep learning workloads. This platform offers on-demand access to high-performance GPUs suitable for intensive training tasks. For this phase, I used a single NVIDIA A100-PCIE-40GB GPU. The server was equipped with a 10-core Intel Xeon CPU, running CUDA version 12.2 and GPU driver version 550.90.07.

Category	**Details**	**Phase**
Platform	AWS	Reasoning Data Generation
	AutoDL	Model Training & Fine-tuning
Instance/GPU	g5.xlarge / NVIDIA A10G	Reasoning Data Generation
	Single NVIDIA A100-PCIE-40GB	Model Training & Fine-tuning
CPU	—	Reasoning Data Generation
	10-core Intel Xeon CPU	Model Training & Fine-tuning
CUDA Version	—	Reasoning Data Generation
	12.2	Model Training & Fine-tuning
Driver	—	Reasoning Data Generation
	550.90.07	Model Training & Fine-tuning

Table 3.3: Computational environments for reasoning data generation and model training.

Chapter 4

Experiments and Results

This chapter presents a set of experiments following the methodologies proposed in the previous chapter. I start by establishing crucial performance baselines, which involves first assessing the zero-shot, out-of-the-box pretrained Qwen models to demonstrate the fundamental need for fine-tuning. Then I introduce a state-of-the-art quality estimation model, COMET-DA, to locate this work within the competitive landscape of the field. Following the establishment of baselines, I conduct an ablation study that compares full-parameter fine-tuning against the parameter-efficient LoRA method, which justifies the use of LoRA in the main training process. Main experiments are highlighted, which directly contrast the performance of the student model finetuned with gold scores (referred to as sft-gold) against the student model finetuned with GPT-generated reasonings (referred to as sft-GPT). Through this structured progression of experiments, I aim to provide a clear view to validate the methodological choices and to demonstrate the impact of incorporating explanatory knowledge in the training process.

4.1 Baselines for Comparison

The purpose of this section is to establish two benchmarks that provide context for all subsequent results. First, a zero-shot test is defined to quantify the performance of the underlying language models without any specific training, in order to prove the necessity of fine-tuning. Second, a state-of-the-art model is introduced to provide a high-performance reference point, which allows to measure the competitiveness and success of the fine-tuned models in this study.

4.1.1 Zero-Shot Performance of Qwen Models

The objective of this initial experiment is to closely evaluate the zero-shot performance of the official pretrained Qwen models, specifically Qwen3-0.6B and Qwen3-1.7B. This experiment is designed to establish a potentially lower-bound performance benchmark. By quantifying how the models perform without any task-specific training, I can validate the hypothesis that quality estimation capability must be learned through targeted training. The models used here are the original, full-precision versions released by the developers, not the 4-bit quantized versions utilized in later fine-tuning experiments.

To interact with the models, I employed a structured, prompt-based approach. The input was formatted using the AutoTokenizer's built-in chat template feature, which organizes the conversation into a standardized format that the model is pretrained to understand. The prompt was composed of a system message, which sets the context and defines the model's persona and task, and a user message, which provides the specific data for evaluation.

To manage the model's generation process, the *max_new_tokens* parameter was set to 10. This value provides sufficient length for generating a numerical score with up to four decimal places, preventing the model from producing extra text. This entire evaluation was conducted in a strict zero-shot setting, meaning no examples of correctly scored pairs were provided. The models relied solely on the pretrained knowledge and the instructions in the system prompt. The evaluation was run on the complete test set of 14,477 sentence pairs. The predicted numerical scores were precisely extracted using regular expressions for comparison against the gold standard Z_{mean} scores.

The zero-shot performances of both out-of-the-box Qwen models are presented together with the baseline model in the table4.1.

The experimental results demonstrate the inadequacy of the fresh out-of-the-box pretrained models for this complex QE task. The Qwen3-0.6B model yielded a Pearson correlation coefficient of a mere 0.0076, a value very close to zero which indicates the absence of a meaningful statistical relationship with the human-annotated Z_{mean} scores. The outputs reveal that the model predicted the score ”-2.0” for the vast majority of the 14,477 test samples. This behavior suggests that the model was incapable of scoring effectively. Instead, it adopted a simplistic method of echoing the numerical value that is shown in the prompt.

The larger Qwen1.7B model showed a marginal improvement, as expected due to its increased parameter count. Its Pearson correlation rose to 0.1906, suggesting a weak positive correlation, but this is still far below acceptable. Mean Absolute Error (MAE) quantifies the average size of the errors between predicted and true values. The Qwen1.7B model's MAE of 0.7940 is less than half that of the 0.6B model's 1.7388, indicating that the predictions were closer to the true scores. Mean Squared Error (MSE) emphasizes larger deviations by squaring the difference between predicted and true values. The dramatic drop in MSE from 3.7506 to 1.1254 shows that the larger model is more stable and made fewer severe errors. This is further reflected in the Root Mean Squared Error (RMSE), where the 0.6B model scored 1.9366, not comparable with the 1.7B's 1.0608.

These findings confirm that the general capability of the pretrained models is insufficient to specialized QE task. The task requires both understanding the various languages and performing a complex process of comparison, error identification, and mapping that qualitative judgment to a precise score in a required range. The zero-shot results prove that this ability must be taught through further fine-tuning.

4.1.2 State-of-the-Art Baseline: wmt22-cometkiwi-da

As a strong reference-free QE model, wmt22-cometkiwi-da is able to predict a quality score based solely on the source and translated sentences, which perfectly aligns with the task definition. This specific version was trained on data from the WMT22 shared task, which has a significant overlap in data sources and language pairs with the training dataset of this task, making it a suitable and challenging baseline. I ran the official COMETKiwi-DA model on the entire test set, generating a quality score for each sample. The predicted scores fall within the range of [0, 1]. They were then compared against the Z_{mean} gold scores to calculate the four standard performance metrics.

The performance of the COMETKiwi-DA model is presented in the following table, juxtaposed with the zero-shot baseline results to highlight the performance gap between an untuned SLM and a specialized SOTA model.

Model	Pearson r	MAE	MSE	RMSE
Fresh Qwen3-0.6B	0.0076	1.7388	3.7506	1.9366
Fresh Qwen3-1.7B	0.1906	0.7940	1.1254	1.0608
COMETKiwi-DA	**0.5984**	0.8621	1.2633	1.1240

Table 4.1: Performance comparison across three models

As expected, the COMETKiwi-DA model demonstrates a very strong performance, achieving a Pearson correlation of 0.5984 which signifies a strong, positive linear relationship with human judgments. An interesting nuance appears in the other metrics: COMET's MAE (0.8621) and MSE (1.2633) are slightly higher than those of the zero-shot Qwen3-1.7B model. This may suggest that the 1.7B model achieved a lower average error by making ”safe”, low-variance predictions, while COMET produces a much wider and more granular range of scores which mirrors the subtle changes in translation qualities. COMET excels at correctly identifying which translations are better or worse than others, which is reflected in the high Pearson r value. Therefore, the performance of COMET establishes a formidable yet clear benchmark for the models developed in the subsequent stages of this research.

4.2 The Impact of Fine-tuning Strategy

In this section, I compare the Full Fine-tuning (FFT) approach with the Parameter-Efficient Fine-tuning (PEFT) using LoRA to provide a clear evaluation of their trade-offs. Both FFT and PEFT are common strategies. By analyzing the performance metrics and resource requirements, I aim to validate the decision to use LoRA as the primary training strategy for the main experiments.

Two parallel experiments were conducted using the Qwen3-0.6B model, with both being trained exclusively on the gold standard Z_{mean} scores.

- **Experiment 1 (FFT):** In this experiment, I fine-tuned all 0.6 billion parameters of the official pretrained Qwen3-0.6B model. This represents the traditional approach where the entire weight matrix of the model is updated during training.
- **Experiment 2 (PEFT with LoRA):** In this experiment, I utilized the Unsloth framework and a 4-bit quantized version of the model. I then applied LoRA to freeze the majority of the model's weights and trained only a small number of injected adapter layers. I expect that this approach is used to avoid catastrophic forgetting of the model's pretrained knowledge, to enhance training stability, and to reduce computational and memory requirements.

I set the rank (r) and *lora_alpha* to 32, acts as a scaling factor for the learned adaptations. The training process utilized *bf16=True*, enabling BFloat16 mixed-precision training to accelerate computation while maintaining stability. The *optim=”paged_adamw_32bit”* setting engaged a memory-efficient variant of the AdamW optimizer, which is effective for quantized models.

The performance and efficiency of the two fine-tuning strategies are directly compared in the following table.

Model	**Pearson r**	**MAE**	**MSE**	**RMSE**	**Training Time**
FFT Qwen3-0.6B	0.2994	0.6166	0.7569	0.8700	∼16 hours
LoRA Qwen3-0.6B	**0.4513**	0.6002	0.6908	0.8311	∼10 hours
COMETKiwi-DA	**0.5984**	0.8621	1.2633	1.1240	

Table 4.2: Performance and training time comparison between full fine-tuning and LoRA fine-tuning

The results show a dramatic improvement of the Pearson r. The fully fine-tuned model's correlation of 0.2994 demonstrates that the model has successfully learned the fundamental relationship between translation quality and the numerical score. However, the model tuned with LoRA exhibits a stronger performance and more economic in resources. The improvement is over 50%. A possible explanation is that LoRA acts as a form of regularization by constraining the number of trainable parameters. LoRA may have prevented the model from overfitting to the 134k examples in the training set and encouraged the model to learn more generalizable features of translation quality. The advantage of using LoRA is quite clear in terms of both performance and computational resources.

In conclusion, this experiment provides a strong justification for the methodology choice. LoRA is not merely a compromise for efficiency; in this study, it proved to be the better method, delivering both performance improvement and a significant reduction in computational cost and training time. Therefore, it was selected as the sole fine-tuning strategy for the main experiments.

4.3 Evaluating Reasoning-Enhanced Distillation

This section presents the core experiment of this thesis, designed to directly answer the primary research question: can the integration of LLM-generated natural language reasoning effectively enhance the student model's ability in machine translation QE task? I hypothesize that providing natural language reasoning as an auxiliary training signal helps the student to learn the scoring mechanism more effectively compared to supervision on only gold Z_{mean} scores. This auxiliary supervision is expected to help the student model to capture more subtle quality indicators, leading to better understanding of the task and better performance on the test set.

To validate the impact of the reasoning supervision, I conducted a direct comparison between two models. Both models share the identical Qwen3-0.6B architecture and were fine-tuned using the same LoRA configuration and hyperparameters as detailed in Section 3.4. The sole distinction between them lies in the training inputs.

- **Student fine-tuned with gold scores (sft-gold):** This model was fine-tuned with LoRA on the gold scores alone. The input for each training example consisted of the (src, mt) pair, and the model was tasked with regressing the corresponding Z_{mean} score.

- **Student fine-tuned with GPT reasonings (sft-GPT):** This model was fine-tuned using the same LoRA strategy but on the augmented dataset. For each training example, the input prompt was enriched to include the (src, mt, reasoning) triplet, where the reasoning was the text generated by GPT-4o-mini. The learning objective remained the same: to regress the Z_{mean} score.

Both models utilized the same system prompt during training, ensuring that the only variable being tested was the presence of the GPT-generated reasoning in the user prompt.

The performance of the two models, along with all previously established baselines, is presented in the following table, giving a direct and contextualized comparison.

Model	**Pearson r**	**MAE**	**MSE**	**RMSE**
Fresh Qwen3-0.6B	0.0076	1.7388	3.7506	1.9366
Fresh Qwen1.7B	0.1906	0.7940	1.1254	1.0608
FFT Qwen3-0.6B	0.2994	0.6166	0.7569	0.8700
LoRA Qwen3-0.6B	0.4513	0.6002	0.6908	0.8311
Qwen3-0.6B with GPT Reasoning	**0.4820**	0.7336	0.9634	0.9815
COMETKiwi-DA	**0.5984**	0.8621	1.2633	1.1240

Table 4.3: Comparison of model performance with different fine-tuning strategies.

The results of the experiments present the impact of reasoning-enhanced distillation.

A direct comparison between the two main models reveals that sft-GPT achieved a higher Pearson r (0.4820) than sft-gold (0.4513). This improvement of 0.0307 is less high than expected but still shows the impact of reasoning supervision. It indicates that the predictions of sft-GPT have a stronger linear relationship with human judgments. It means that the model improved in correctly ranking the quality of translations. This supports the hypothesis that the reasoning supervision helps the model capture the underlying quality indicators within the data more effectively.

However, an interesting trade-off emerges when examining the other error metrics (MAE, MSE, and RMSE). Sft-gold consistently outperformed the sft-GPT on all three of them. The lower MAE (0.6002 vs. 0.7336) of sft-gold suggests that on average, its predictions were numerically closer to the true Z_{mean} scores. The lower MSE and RMSE values further indicate that its predictions were more stable and less prone to large, outlier errors. Despite the better correlational performance of sft-GPT, the other metrices exhibit that the predictions were numerically more deviant from the ground truth. This suggests a potential issue: the model could correctly identify the quality of a translation, but over- or underestimates the magnitude of the quality difference.

Several factors could lead to this outcome. First, there may be a degree of information redundancy between the gold score and the reasoning texts. The Z_{mean} score already implicitly contains all the information about the quality of the translations. While the reasoning makes this information explicit, a significant portion of it might be redundant. Second, the hyperparameter configuration, while optimal for the baseline task, might be less optimal for reasoning supervision, since the reasoning-augmented prompts are much longer. The current learning rate or number of epochs might be insufficient for the model to fully grasp the rich, unstructured information present in the reasoning text. Finally, the reasoning texts might introduce subtle biases or a specific linguistic style from the teacher LLM, which could lead to predictions that hold on to the same biases.

Overall, the experiments successfully demonstrates that incorporating LLM-generated reasoning as an auxiliary signal can improve a student model's ability on QE task, as evidenced by the increase in Pearson correlation. However, it also reveals the trade-off between correlational alignment and absolute error, suggesting that future work could focus on refining the training process to balance the benefits of reasoning while improving numerical precision.

Chapter 5

Error Analysis

In this chapter, I conduct error analysis on two main models: the student model fine-tuned with gold scores (“sft-gold”) and the student model fine-tuned with GPT rationales (“sft-GPT”). The analysis includes two parts: quantitative analysis, examining predictions with respect to gold scores, sentence length, and language pairs; and qualitative analysis, focusing on the most overestimated, underestimated, and disagreed cases between the two models.

5.1 Quantitative Analysis

5.1.1 Pearson r by Z_{mean} scores range

Following the training prompt, the Z_{mean} range of all the 14,477 samples in test set is divided into six different ranges:

1. High Quality in long tail end (≥ 2.0). This range is considered as a long tail end in the whole distribution due to the high scores and few amounts of samples.
2. High Quality in range ($1.0 \sim 2.0$],
3. Medium High Quality ($0.0 \sim 1.0$],
4. Medium Low Quality ($-1.0 \sim 0.0$],
5. Low Quality ($-2.0 \sim -1.0$],
6. Very Low Quality (≤ -2.0). This range is considered as a long tail end in the whole distribution due to the low scores and few amounts of samples.

1. High-quality samples in long tail end

There are 57 samples in this range. Sft-GPT excels with a Pearson r of 0.49, while sft-gold delivers a weak performance with 0.229. However, sft-gold has a RMSE of 1.377 which is significantly better than 2.146 of sft-GPT.

Metric	**sft-gold**	**sft-GPT**
Pearson r	0.229	**0.490**
RMSE	1.377	2.146

Table 5.1: Comparison of sft-gold and sft-GPT in High Quality Long Tail Range

The distribution of the test set follows a roughly normal trend, with most data concentrated in the quality score range of $[-2, 2]$. Outside this range lies a small portion of data with more extreme quality scores, referred to as the long-tail range. In prediction, although the models are only required to produce quality scores within four levels in the $[-2, 2]$ range, it is still possible to calculate the correlation between Z_{mean} and the predicted scores for the same examples in the long-tail region. This also explains why both models show high RMSE values. Due to score compression, the models cannot produce consistent absolute scores across the full range, and can only aim to preserve correlation between the scores. For targets outside the range, the models tend to output values close to 2.0, while the gold scores may be 2.5, 3.0, or higher, leading to large absolute errors. Sft-GPT still achieved a high correlation score, but its accuracy declined further compared to sft-gold.

2. High-quality samples

There are 829 samples in this range.

Metric	**sft-gold**	**sft-GPT**
Pearson r	**0.127**	0.018
RMSE	**0.645**	1.355

Table 5.2: Comparison of sft-gold and sft-GPT in High Quality Range

In this range, the number of samples has significantly improved. sft-gold outperforms sft-GPT on both correlation and absolute error. Give the larger sample size, these estimates are more stable and suggest a reliable advantage for models finetuned with gold scores. A possible interpretation is that direct supervision with human-annotated Z_{mean} scores provides better numerical calibration within the common quality range, while supervision with GPT-reasoning tends to improve relative ranking ability, but may also introduce bias in absolute score values.

3. Medium High Quality samples

This is the largest group with 7,550 samples.

Metric	**sft-gold**	**sft-GPT**
Pearson r	**0.348**	0.180
RMSE	**0.494**	0.934

Table 5.3: Comparison of sft-gold and sft-GPT in Medium High Quality Range

The Medium-High quality range has the biggest sample size in all six ranges. It's clear that sft-gold still outperforms sft-GPT in both correlation and absolute accuracy. Given the large sample size, this performance gap is highly reliable. It also suggests that the GPT-reasoning supervision for Model may work well in extreme cases, but does not generalize as well as expected in the range where accurate predictions are important.

4. Medium Low Quality samples

This is the second large group with 4,176 samples. In this range, sft-GPT shows a clear advantage in correlation, but slightly worse in accuracy than sft-gold.

Metric	sft-gold	sft-GPT
Pearson r	0.096	**0.259**
RMSE	**0.812**	0.901

Table 5.4: Comparison of sft-gold and sft-GPT in Medium Low Quality Range

5. Low Quality samples

There are 1,550 samples in this range. sft-gold and sft-GPT show similar rank correlation, but sft-GPT achieves much lower RMSE. This pattern suggests that both models roughly agree on the ranking order, while sft-GPT is better calibrated in accuracy for poor translations.

Metric	sft-gold	sft-GPT
Pearson r	**0.196**	0.190
RMSE	1.435	**0.911**

Table 5.5: Comparison of sft-gold and sft-GPT in Low Quality Range

6. Very Low Quality samples

There are 310 samples in this long tail range. Both models show negative correlations with gold Z_{mean} scores in this range, which indicates that the ranking is not reliable in this extreme case. The slight advantage of sft-GPT in RMSE still shows a better numeric accuracy although the ranking fails. Negative correlation in this range is expected. The scores cluster near the bottom line, even minor difference in prediction could lead to a shift in rankings. However, the better RMSE of sft-GPT suggests that its outputs remain numerically closer to gold scores.

Metric	sft-gold	sft-GPT
Pearson r	-0.199	-0.195
Same (with negative correlation)		
RMSE	2.442	**1.701**

Table 5.6: Comparison of sft-gold and sft-GPT in Very Low Quality Range

With all the six quality ranges together, it is clear that the varying performance of the two models reflects distinct strengths and weaknesses according to their training supervision. Sft-gold, finetuned solely on gold Z_{mean} scores, consistently achieves better accuracy and correlation in medium to high-quality ranges (Ranges 2 and 3). It shows a strong alignment with human judgments where translation quality is moderate to good. It suggests that direct supervision with gold scores effectively guides the model to capture nuanced quality distinctions in these ranges.

Conversely, sft-GPT, which incorporates GPT-generated reasoning during training, shows advantages in very high and low-quality extremes (Ranges 1, 4, 5, and 6). The higher correlation in the extreme high-quality range indicates improved sensitivity to top-tier translations, which possibly comes from GPT's richer reasoning signals in the training inputs. Similarly, in lower quality ranges, sft-GPT achieves lower RMSE, suggesting better numerical precision when identifying poor translations, possibly because the reasoning data helps the model recognize nuanced error patterns missed by sft-gold. However, sft-GPT' s overall precision suffers in the mid-quality ranges. The higher RMSE and lower correlation indicate a trade-off, where reasoning guidance might introduce noise. Another possibility is that the GPT reasonings gave overlapping information which also appeared in the gold Z_{mean} scores, leading to an overfitting on the model's judgement.

The sample size also impacts the results. In extreme ranges where scores are contracted, both the models tend to show less robustness in accuracy. In the mid-ranges where sample sizes are large, the influence of supervisions are fully demonstrated. This highlights again that gold score supervision ensures reliable performance in common quality ranges, while GPT reasoning helps the detection of edge cases and extreme quality levels.

5.1.2 Pearson r by Sentence Length

In the whole test set, 33% sentences are shorter than 12 words, and 66% sentences are shorter than 18 words. The longest sentence consists of 127 words, while the shortest sentence has only 1 word. I define sentences that have less than 12 words as short sentences, from 12 to 18 words as medium sentences, longer than 18 words as long sentences.

Sentence Length	Samples	Pearson r	RMSE
1–12	5548	**0.4964**	0.8909
12–18	4168	0.4289	0.8513
18+	4721	0.3952	0.7347

Table 5.7: sft-gold performance across sentence lengths

Sentence Length	Samples	Pearson r	RMSE
1–12	5550	**0.5173**	1.0463
12–18	4181	0.4635	1.0042
18+	4742	0.4526	0.8773

Table 5.8: sft-GPT performance across sentence lengths

Sentence Length	Pearson r Difference	RMSE Difference
1-12	+0.0209	-0.1554
12-18	+0.0346	-0.1529
18+	+0.0574	-0.1426

Table 5.9: Differences in Pearson r and RMSE by sentence length group.

The performance of both models varies across different sentence length groups. For short sentences, both models achieve their highest Pearson correlation scores, with sft-GPT slightly outperforming sft-gold. This suggests that short sentences, typically simpler in meaning and less complex in syntax, are easier for both models to evaluate. Sft-gold maintains its advantage in RMSE, showing that it's more precise in numerical predictions. In the medium-length group, a similar pattern can be observed. Sft-GPT again achieves higher correlation and sft-gold better numerical precision. For long sentences, sft-GPT gains most relative correlation than sft-gold. The GPT reasoning in sft-GPT's supervision may have helped in better handling the complexity and nuanced errors that are common in longer texts. Although the RMSE of sft-GPT is still higher than sft-gold's, the gap has narrowed compared to that in shorter sentences. Sft-gold gains its best RMSE in the long sentence group, suggesting its numeric predictions are more consistent for longer inputs.

Looking at table **??**, it's clear that sft-GPT improves Pearson correlation scores across all lengths, increasing from short sentences to long ones. Sft-gold consistently achieves lower RMSE by a margin of approximately 0.14 to 0.16, pointing to better score precision overall.

Overall, the integration of GPT reasoning in the training of sft-GPT has been proven effective in improving the correlation of the model's predictions with human judgements, especially when sentences get longer and more complex. In contrast, this come along with the degradation of accuracy, shown by the increasing RMSE. These complementary behaviors suggest that combining reasoning-based supervision with direct score training could further improve MT quality estimation across varying sentence lengths.

5.1.3 Pearson r by Language Pairs

Below are the two models's Pearson r correlation scores across language pairs, listing from high performance to low. The sample size to each language pair may have a slight difference after the NaN values are dropped. There are multiples reasons why the performance varies, from linguistic similarity of the source/target language to English, the morphological complexity, to the dataset size.

1. High Performance Group ($r > 0.5$) In this group, there are 5 language pairs. Sft-gold presents 6,219 samples occupying 43.1% of the whole test set, with an average Pearson r 0.586. Sft-GPT presents 5,647 samples, occupying 38.8% of test set, with an average Pearson r 0.625.

Language Pair	Size Bucket	sft-gold r	sft-GPT r
ro-en	1000–2000	0.693	0.720
ru-en	1000–2000	0.682	0.755
ne-en	1000–2000	0.537	–
et-en	500–1000	0.509	0.594
en-zh	1000–2000	0.507	–
en-ta	500–1000	–	0.538
si-en	1000–2000	–	0.520

Table 5.10: Comparison of Pearson r for high-performance language pairs. Sample sizes are binned into ranges; exact counts are given in Appendix 1.

Looking at sft-gold first: in the High Performance group ($r > 0.5$), ro–en achieves the highest correlation (0.693), although the sample size is not the biggest(1,350). Romanian' s relatively close syntax and vocabulary to English has likely facilitated the translation and reduced the ambiguity, which also lead to strong correlation from the predicted scores and the gold human annotations. Ru–en follows closely (0.682) and has the largest sample size in this group (1,433). The morphological complexity of Russian could bring difficulty in translation, but here a stable correlation is established, which shows that a large dataset size could digest the potential difficulties and add in better generalization from the models.

For sft-GPT, the Pearson r correlation scores for the highest two language pairs (ru-en and ro-en) have largely improved, indicating a consistent boost in ranking accuracy for linguistically closer pairs. Et–en (0.594) and en–ta (0.538) both perform well with small sample sizes (709 and 705, respectively), which points to the model' s ability to generalize effectively in limited-data scenarios. For et-en and si-en, gains are notable. Si-en and en-ta pairs are in medium performance group in sft-gold's predictions, but they appear interestingly in the high performance group of sft-GPT. This may suggest that sft-GPT is better at handling languages of low-resource and of morphological complexity than sft-gold. Both the two models are very good at X-to-English scenarios, showing an obvious language direction preference.

2. Medium Performance Group ($0.3 \leq r \leq 0.5$) In this group, there are 5 language pairs. Sft-gold presents 6,112 samples occupying 42.3% of the whole test set, with an average Pearson r 0.367. Sft-GPT presents 6,719 samples, occupying 46.2% of test set, with an average Pearson r 0.470.

Language Pair	Size Bucket	sft-gold r	sft-GPT r
en-ta	500–1000	0.426	–
en-gu	500–1000	0.376	0.484
en-mr	2000+	0.367	0.494
si-en	1000–2000	0.350	–
en-hi	500–1000	0.316	0.482
en-zh	1000–2000	–	0.445
ne-en	1000–2000	–	0.443

Table 5.11: Comparison of Pearson r for medium-performance language pairs. Sample sizes are binned into ranges; exact counts are given in Appendix 2.

The Medium Performance Group ($0.3 \leq r \leq 0.5$) displays greater diversity in sample sizes and correlations. En–ta (0.426) ranks highest in sft-gold's performance, showing again that even morphologically rich languages with low-resource status can achieve moderate correlation on smaller sample size (705). Si–en (0.350) deliver moderate results, with Sinhala' s large sample size (1,425) suggesting that large sample size alone cannot always guarantee high correlation. The largest dataset overall, en–mr(2,652), yields a correlation of 0.367—higher than en–hi (0.316), but still below the high-performance threshold. Sft-GPT's performance shows a significant improvement in Indic language pairs, with en-hi raised 52.5% and si-en raised 48.6%. The performance on en-mr has also raised drastically, given the largest sample size in the whole dataset. En-zh has degraded greatly, falling from the high performance category to medium-low. This suggests that improvements are not evenly distributed across languages, and the morphological characteristics of the Chinese language may have hindered the adaptation process. Overall, sft-GPT' s medium-range performance reflects a strong net gain, but also exposes language-specific weaknesses that would need targeted strategies.

3. Low Performance Group ($r < 0.3$) In this group, there are 2 language pairs. Sft-gold presents 2,106 samples occupying 14.6% of the whole test set, with an average Pearson r 0.081. Sft-GPT presents 2,107 samples, occupying 14.5% of test set, with an average Pearson r 0.155.

Language Pair	Size Bucket	sft-gold r	sft-GPT r
en-de	1000–2000	0.082	0.135
en-te	500–1000	0.080	0.175

Table 5.12: Comparison of Pearson r for low-performance language pairs. Sample sizes are binned into ranges; exact counts are given in Appendix 3.

The language pair en-de surprisingly shows in the low performance category in both models's performance. For German, it is obvious that the model struggles with high-resource yet structurally complex language, this may be due to nuanced word order variation, compounding morphology, and a relatively high baseline translation quality that reduces score variability. Telugu (en–te), like other Dravidian languages, poses additional difficulties due to its rich inflectional morphology, complex script, and resource rarety. For sft-GPT, only marginal gains compared to sft-gold are delivered, with Pearson r values of 0.175 for en-te and 0.135 for en-de. Although both language pairs show slight improvements, their correlation scores remain well below 0.3, indicating that sft-GPT still fails to capture consistent quality signals in these cases. Although avoiding regression in low range, it also fails to deliver meaningful breakthroughs, accentuating the need for language-specific model adaptations.

Overall, there are 10 language pairs where sft-GPT gives better performance, 2 language pairs where sft-gold excels. Sft-GPT shows significant improvement on ”weak language pairs”: the improvement is even greater for language pairs that originally performed poorly. A strong ”language family effect” is observed, where Indic languages benefit the most, while Russian and Romanian languages continue to maintain their advantage and further improve, and German fail despite the similarity to English. This shows that sft-GPT has achieved better cross-lingual generalization capabilities through GPT reasoning training, especially showing obvious advantages when dealing with language pairs where sft-gold performs poorly.

5.2 Qualitative Analysis by MQM Category

In this part, for each model, 20 cases of over-predictions, 20 cases of under-predictions, and 20 most disagreed cases between the two models are picked out and analyzed. Representative errors by MQM category are examined to provide insight into common weaknesses of the performance. Accuracy issues account for a large portion of severe errors. They consist of terminology and named entity mistranslations, leaving a big semantic impact in domain-specific contexts, such as religious texts. Fluency errors such as grammar and syntax mistakes reflect insufficient target-language modeling, especially in morphologically rich languages. Style errors indicate register inconsistency and lexical redundancy in translations, which make professional or academic texts less formal. From these findings, we can see the importance to improve domain-specific vocabulary coverage, to strengthen the fluency in target language, and to enhance cultural adaptation to target culture while preserving the meaning from the source language.

5.2.1 1. Most Over-predicted Cases

Examples of the over-predicted cases of sft-gold can be found in Sft-gold's performance in Over-predicted cases. In sft-gold's predictions, the cross-language analysis reveals that the *en-de* pair exhibits recurring weaknesses in 11 out of 20 examined cases, including insufficient attention to terminological precision in German, underestimation of grammatical complexity, tolerance of English–German mixed expressions, and neglect of register requirements in academic or formal contexts. For other language pairs, *en-hi* shows excessive tolerance toward terminological inaccuracies and syntactic deviations, *en-mr* demonstrates comparatively stricter adherence to semantic accuracy, and *en-ta* displays inadequate detection of semantic omissions.

Examples of the over-predicted cases of sft-GPT can be found in Sft-GPT's performance in Over-predicted cases. The GPT reasoning supervision for sft-GPT shows a dual-side impact. On the positive side, the model demonstrates accurate identification of most linguistic issues, provides detailed grammatical and semantic analyses, and considers multiple dimensions such as accuracy, fluency, and register. On the negative side, it applies overly lenient scoring standards, often assigning high scores despite recognizing clear errors; it overlooks domain-specific precision requirements; and it shows inconsistency in scoring the same type of error across different contexts.

Over-predictions appears due to multiple reasons. For sft-gold, over-predictions appears to come from a surface-level evaluation approach, focusing on lexical correspondence and formatting, blind spots in semantic consistency detection, and insufficient sensitivity to target-language specificity. For sft-GPT, the causes include an overly tolerant ”understanding is sufficient” evaluation logic, misweighted importance of domain-specific precision, and inadequate sensitivity to quality requirements across different application contexts.

On linguistic-Level, the two models follows different philosophies of translation quality evaluation: sft-gold leans toward a *form-based* evaluation, while sft-GPT favors *functional equivalence* but applies it with excessive leniency. On the other hand, sft-GPT shows stronger capabilities in error categorization, but with consistent bias in assessing error severity. Sft-GPT seems to assume a target user who tolerates minor errors, which is unsuitable for professional translation scenarios.

Below is a table showing the difference of both models in quality estimation. Detailed cases can be found in the Appendix part.

Aspect	Sft-gold	Sft-GPT	Difference Analysis
Technical Terms	Under-recognition	Over-tolerance	Sft-GPT knows issues but rates too leniently
Grammar Errors	Surface-level evaluation	Theoretical understanding but lenient in practice	Sft-GPT understands deeper but applies lower standards
Semantic Consistency	Serious blind spots	Relatively sensitive	Sft-GPT clearly outperforms sft-gold

Table 5.13: Differences in Error Severity Recognition

5.2.2 2. Most Under-predicted Cases

Examples of the under-predictions cases of sft-gold can be found in Sft-gold’s performance in Under-predicted cases. From the picked cases, it’s clear that sft-gold demonstrates a consistent tendency to severely under-predicted translation quality in cases where the translations are relatively reasonable but contain minor imperfections. The model tends to over-penalize small lexical or grammatical deviations, such as missing articles or mixed-language proper nouns in English-German pairs, treating them as major errors. It also marginalize language pairs with large structural differences or typological distance such as Sinhala-English and Nepali-English, particularly penalizing necessary word order changes or code-switching phenomena. Identical errors receive different ratings depending on these language pairs, revealing a language hierarchy bias. Cultural and contextual factors are frequently misunderstood or ignored, especially in religious or historical texts. The challenges of translating classical historical documents are under-predicted. These tendencies highlight sft-gold's limitations in balancing surface-level fluency, semantic fidelity, and culturally informed interpretation across diverse language pairs and text genres.

The over-prediction and under-prediction cases of sft-gold show a contradictory evaluation philosophy. In over-predicted cases, sft-gold tolerates clear semantic deviations and terminology errors, yet in under-predicted cases, it harshly penalizes minor grammatical omissions and cultural adaptation issues. The inconsistency is also reflected in uneven treatment across language pairs. For German as a target language, sft-gold applies more lenient standards. However, for low-resource languages as source language and English as target language such as Sinhalese-English and Nepali-English, it imposes stricter requirements on the target language. This explains why language pairs such as Si-en and Ne-en performed poorly in the previous analysis, which may be partly due to evaluation bias rather than problems with translation quality itself.

Examples of the under-predicted cases of sft-gold can be found in Sft-GPT's performance in Under-predicted cases. With the help of GPT-generated reasonings, sft-GPT demonstrates strong performance in the evaluation of under-predicted cases, particularly in its accuracy of error identification. It gives good performance at precisely classifying both semantic and grammatical errors, such as content distortion, concept confusion, verb tense, and syntactic structure. For special domains, sft-GPT strictly assesses the accuracy of technical and scientific terminology. The reasoning supervision brings deep linguistic analysis with concrete examples of mistranslated terms. It examines the impact of errors on overall comprehension, and also account for the grammatical norms of the target language. Semantically, sft-GPT applies appropriately strict standards according to text style and register. It demands high precision in scientific and technical texts, emphasizes accuracy in historical and political content, and shows careful scrutiny in handling proper nouns. As for bias patterns, sft-GPT's accentuation on Nepali-English cases (10 out of 20) is based on real errors, while sft-gold shows overstrict criteria towards south Asian languages. Specific patterns of the two models by language pairs can be found in Language Pair Specific Patterns.

Sft-GPT has several clear advantages. It can accurately tell the difference between small problems and serious errors. Its judgments are based on the real quality of the language, not just surface impressions. The model also keeps high standards for technical and scientific texts, where accuracy is highly required.

Below is a table showing the difference in both models's evaluation performance.

Aspect	Sft-gold	Sft-GPT
Language Bias	Surface-level evaluation	Based on actual error severity
Assessment Basis	Fluency prioritized, ignores information completeness	Semantic accuracy prioritized
Cultural Sensitivity	Lack of understanding, over-punishment	Relatively fair, quality-based
Professional Standards	Inconsistent, biased	Strict but reasonable

Table 5.14: Differences on Under-prediction Reasons

5.2.3 3. Most Disagreed Cases: A glance on GPT's bias

Although the models are prompted with the same instrction, the 20 cases of largest disagreement reveal several key divergences between two models. From the aspect of evaluation philosophy, sft-GPT shows more strict standard towards severe semantic inaccuracies, while sft-gold appears to put more weight on surface-level linguistic coherence, assigning relatively lenient scores on the same case. From the aspect of error-type sensitivity, the two models respond differently to critical errors such as complete semantic reversals, mistranslations of domain-specific terminology, and inaccuracies in numerical information, directly impacting their reliability in practical applications. In terms of language pairs, the disagreements are concentrated in English–South Asian and certain European pairs, potentially reflecting differences in training data coverage and in the capacity of the models to handle linguistic complexity.

The disagreement range spans from 3.08 to 3.19 points, with sft-gold assigning nuanced positive predictions (0.536 to 1.182) and sft-GPT assigning extremely negative predictions (-2.6 to -2) across all cases, indicating sharp difference of the evaluation philosophy. The difference patterns are further analyzed in the following cases:

1. **Complete semantic reversal**
   Case 1 (largest disagreement 3.1892):

   - Language Pair: English-Guajarati (en-gu)
   - Error: BJP accuses Congress → Congress accuses BJP
   - Gold Z_{mean} score: -2.655
   - Sft-gold prediction: 0.888, **Sft-GPT prediction: -2.301**
   - Analysis: sft-GPT correctly identified this semantically catastrophic error, while sft-gold, possibly overemphasizing surface fluency, failed to adequately penalize the severity of the semantic distortions. This reveals a potential shortage in handling errors involving logical relations.

2. **Mistranslation of cultural/historical terms**
   Case 2 (Disagreement 3.1853):

   - Language Pair: Sinhala-English (si-en)
   - Error: ”mixed farming” → ”mixed responsibility”
   - Gold Z_{mean} score: 0.781
   - **Sft-gold prediction: 1.182**, sft-GPT prediction: -2.003
   - GPT explanation: Score: -2. The translation misrenders “mixed farming/crops” as “mixed responsibilities,” uses the awkward passive “was referred to” instead of a phrase like “was directed toward” or “aimed at,” and thus entirely distorts the original idea of agricultural intent in the American colonies, making it essentially incomprehensible.

- Analysis: From GPT explanation, it's clear that sft-GPT put a lot emphasis on the mistranslation of key terms, thus give a low score, while ignoring the relative integrity of the syntactic structure.

Case 3 (Disagreement 3.118):

- Language Pair: Romanian-English (ro-en)
- Error: "armistice" → "notification"
- Gold Z_{mean} score: 0.915
- **Sft-gold prediction: 1.078**, sft-GPT prediction: -2.04
- GPT explanation: Score: -2. The translation commits a critical mistranslation by rendering amistițiul ("the armistice") as "the notification," which completely alters the meaning of the sentence; aside from this semantic disaster, other elements (e.g., "same railway wagon") are passable but the core information is lost, making the translation essentially incomprehensible in context.
- Analysis: sft-GPT overpunished the mistranslation of terminology, overlooked the correct translated parts.

3. **Mistranslation of numerical information**
   Case 4 (Disagreement 3.1347):

- Language Pair: Nepali-English (ne-en)
- Error: "Mahendranagar has a population of over one hundred thousand" → "The majority of the population belongs to mahendranagar rather than aelakh."
- Gold Z_{mean} score: -0.657
- Sft-gold prediction: 1.135, **sft-GPT prediction: -2**
- GPT explanation: Score: -2. The translation completely misrenders the meaning: "there are more than 100,000 people" is turned into a meaningless comparison "belongs to mahendranagar rather than aelakh" , misreading "one lakh" as "aelakh" and substituting "majority" and "belongs" in place of "population size," thus failing to convey the original fact.
- Analysis: The translation converted the numerical information to a meaningless comparison. Sft-GPT's prediction is closer to the Z_{mean} score.

Patterns of high-disagreement language pairs reveal distinct tendencies. For English–South Asian pairs (Gujarati, Sinhala, Nepali), divergences are concentrated in the evaluation of word order, case marking, and complex grammatical structures, with sft-GPT potentially imposing heavier penalties for grammar-specific errors of these languages. For European language pairs such as Romanian–English, disagreements are shown in the aspect of historical terminology and domain-specific vocabulary, where sft-GPT more accurately captures the severity of terminological mistranslations. Interestingly, despite the overall modest performance of English–German, none of the most extreme disagreement cases involve this pair, which may suggest that its translation errors more often fall into certain categories such as article misuse or direct lexical transfers between English and German.

We can see that in the trade-off between fluency and accuracy, sft-gold exhibits a preference for fluency, which reflects the impact of supervision with gold Z_{mean} scores. Sft-GPT, by contrast, applies stricter penalties for accuracy-related errors, a tendency that may come from the emphasis on "reasonableness" in the GPT reasoning text used for its training.

The supervision from GPT reasoning shows a directional bias in its notable sensitivity to mistranslations. Terms such as "catastrophic," "completely distorting," "incomprehensible," and "severe mistranslation" appear frequently, assigning inappropriate weight to the wrongly translated parts and leading the model to almost entirely overlook the big portion of correct aspects within the same sentence. Interestingly, although the prompts in GPT' s reasoning generation did not request scoring, many reasoning statements spontaneously included a "-2" score for the translation, the student model then reproduced the score from the teacher with fidelity. This reveals a potential risk of relying exclusively on GPT: hallucinations embedded in its reasoning text could hinder the evaluation model' s ability to fairly assess translation quality.

The characteristics of the two models have direct impacts for translation quality evaluation system development. Sft-GPT' s evaluation style, with its strong emphasis on the accuracy of key terminology, supports the creation of high-precision translation systems. Sft-gold' s evaluation style, by contrast, puts greater weight on fluency and gives comparatively lighter penalties for accuracy-related errors. In practical applications, sft-GPT' s evaluations may be more reliable for key-information translations in domains such as law and medicine, where terminology plays a critical role. For creative text translation quality evaluation, it's better to balance the advantages of both models.

Chapter 6

Discussion

This work evaluated whether Qwen3-0.6B can be distilled for MTQE task using both human gold scores and GPT-generated reasoning. Reasoning-enhanced training improved rank correlation in extreme cases, while gold-score–only training achieved better calibration and lower errors in common ranges. The series of experiments show that the student model finetuned with GPT reasoning has a slight advantage of 0.03 in Pearson r over the student model finetuned with Z_{mean} scores, while showing some internalized biases from the teacher model.

6.1 Advantages and Limitation

Sft-gold exhibits consistently lower error metrics (MAE, MSE, RMSE) across most quality ranges, particularly in the medium-to-high quality segments where the majority of data is concentrated. This suggests that direct supervision with human-annotated Z_{mean} scores is more effective in stablize the model' s absolute predictions and yielding more precise numerical predictions. Overall, the Pearson r score still remains competitive.

Sft-GPT achieves higher Pearson correlation in the overall test set and in several specific scenarios, including very high and very low quality extremes, as well as longer sentences. These improvements indicate that incorporating GPT-generated rationales helps the student model better capture the relative ranking of translations, especially in edge cases. However, this improvement in rank correlation comes together with the cost of reduced numerical precision, as shown in the higher absolute errors. This trade-off suggests that while reasoning provides richer semantic context, it may also introduce variance in magnitude estimation, potentially due to stylistic or evaluative biases present in the reasoning text.

6.2 Potential Biases of GPT as a teacher

The GPT-generated reasoning introduces potential biases not present in human annotations. GPT tends to exhibit systematic tendencies in its explanations, such as overemphasizing certain error categories (e.g., precision on certain terms over fluency), adopting a consistent strict standard across certain language pairs, regardless of cultural or linguistic backgrounds. Such biases can influence the student' s QE quality.

Shaped by its training data and prompt context, GPT reasonings may not always align with the criteria used in human Direct Assessment (DA). While DA scores represent aggregated human judgments that represent inter-annotator variability, GPT' s outputs are largely impacted by the prompt, lacking the diversity of human perspectives. This difference could lead to model behavior that correlates well with human rankings but deviates in magnitude or error-type sensitivity. As shown in the chapter of Error analysis, sft-GPT often ranks translations correctly in extreme cases but mismatch their severity compared to human scores.

6.3 Balancing the two teachers

The findings suggest that both the human annotator and GPT as teachers contribute in finetuning the student. Human gold scores provide reliable quantitative targets, while GPT rationales offer qualitative, interpretable context to enhance ranking performance and edge-case sensitivity. A balanced approach could involve multi-task learning, where the student is trained to optimize both correlation with gold scores and alignment with GPT-generated rationales.

Staged distillation could be another possibility. First train the student exclusively on human gold scores to establish precise numerical predictions, then conduct a secondary fine-tuning phase incorporating GPT reasoning as auxiliary data. This phased approach may prevent the variance deviation observed in sft-GPT while retaining its benefits in ranking performance.

6.4 Reproducibility

This study uses a publicly available WMT Direct Assessment datasets (2021–2023). It employs an open-source student model (Qwen3-0.6B). The training process is based on open-source framework, such as Hugging Face Transformers, TRL, and Unsloth to ensure equivalent configurations can be reproduced by other researchers. The training configurations are kept transparent. A careful documentation of hyperparameters, LoRA settings, and hardware specifications is presented in Chapter 4. The experiment results are evaluated using standard metrices (Pearson r, MAE, MSE, RMSE).

The closed source component of this study involves GPT-4o-mini reasoning generation. The outputs are deterministic with fixed prompt and temperature, but availability depends on API access.

Reasoning dataset generated in this work is shared on GitHub to ensure potential inputs for future experiments.

6.5 Future work

From a training strategy perspective, future optimization could explore:

- **Curriculum learning**: introduce reasoning supervision progressively, starting with high-quality gold-score predictions before exposing the model to additional rationales.

- **Specialized prompting**: tailor GPT' s reasoning prompts by language pair, domain, or known error patterns to minimize irrelevant or biased explanations.

- **Hybrid loss functions**: combine regression loss on gold scores with ranking or contrastive losses derived from GPT judgments, in order to ensure both magnitude accuracy and ranking fidelity.
- **Data stratification**: apply reasoning augmentation selectively to underrepresented quality ranges or sentence structures to maximize the utility of GPT supervision where human data is less informative.

Chapter 7

Conclusion

This work examined whether LLM-generated reasoning can enhance knowledge distillation for MTQE. Using Qwen3-0.6B as the student, two supervision settings were compared: gold scores only and gold scores plus GPT-4o-mini rationales. Across WMT 2021–2023 datasets, the reasoning-enhanced model achieved higher Pearson correlation on the whole test set, particularly in very high and very low quality ranges and for longer sentences, suggesting improved ranking fidelity and sensitivity to edge cases. The gold-score-only model consistently produced lower MAE, MSE, and RMSE in medium-to-high quality ranges, indicating better numerical calibration. These patterns highlight the complementary nature of human and GPT supervision: human scores provide stable quantitative targets, while GPT reasoning offers qualitative context that improves relative ranking but can increase variance in score magnitude.

The study also identified limitations, including potential bias in GPT ' s evaluations, language coverage constraints, and the absence of extensive hyperparameter tuning for the reasoning-enhanced setup. Nevertheless, the findings suggest multiple optimization avenues: staged or selective application of reasoning supervision, curriculum learning, hybrid loss functions, and adaptive prompting tailored to language and domain. By strategically integrating human and LLM teachers, it is possible to produce QE models that are both computationally efficient and competitively accurate, paving the way for practical deployment in multilingual, resource-constrained translation workflows.

Appendix

A Sft-gold's performance in Over-predicted cases

The cases involve 6 language pairs, namely en-de (9 examples), en-mr (4 examples), en-hi, en-zh, and ru-en (2 examples), en-gu (1 example).

1.1 Accuracy Errors:

Terminology

Case 1 (en-de, diff: 7.98):

Source: *"Structural Engineering"*

Translation: *"Structure Engineering"*

Error: The technical term "structural" was mistranslated as noun "structure", altering both part of speech and meaning.

Impact: Precision of engineering terminology is critical for professional documentation.

Case 2 (en-de, diff: 6.96):

Source: *"The Australian National University"*

Translation: *"Die Australian National University"*

Error: The institutional name was partially Germanized, leading to inconsistency; it should be fully translated.

Named Entities

Case 3 (en-de, diff: 7.48):

Source: *"fellows"* (scholars/researchers)

Translation: *"Fellows"* (capitalized only)

Error: The professional term was left untranslated, with only formatting changes applied.

Semantic Distortion

Case 4 (en-mr, diff: 4.39):

Source: A complex sentence about *ryot* (farmer) and the Commission.

Translation: Unrelated Marathi text about war.

Error: Completely incorrect semantic correspondence; translation is entirely off-topic.

1.2 Fluency Errors

Grammar

Case 5 (en-de, diff: 4.95):

Source: *"lost the August 8 Democratic primary"*

Translation: *"verlor den 8. August Democratic Primär"*

Error: Incorrect German grammar; "Democratic Primär" should be "demokratische Vorwahl".

Syntax

Case 6 (en-de, diff: 3.82):

Source: *"Wood defied tradition and ran for a second term"*

Translation: *"Holz trotzte der Tradition und lief für eine zweite Amtszeit"*

Error: "lief für" is not the correct German expression for "to run for" (in elections).

Consistency

Case 7 (en-mr, diff: 3.60):

Source: *"There is a protuberance on the head and open eyes"*

Translation: (Here Marathi sentence is translated into English) *"There is a lung on the head and on the head"*

Error: Redundant repetition of "on the head" without variation.

Case 8 (en-hi, diff: 3.64):

Source: *"It was everything I thought Bohemia probably was"*

Translation: (Here Hindi sentence is translated into English)Simplified to *"I was just thinking about Bohemia"*.

Error: Loss of reflective tone and complex emotional nuance present in the source.

B Sft-GPT's performance in Over-predicted cases

The cases involve 6 language pairs, namely en-de (11 examples), en-hi (2 examples), en-mr, en-zh, en-gu and en-ta (1 example each).

2.1 Minor Error Over-tolerance

Overrating Identical Text

Case 9 (en-de, diff: 7.82, predicted score: 0.9848):

Source: "Faerie Apocalypse, Jason Franks, IFWG Publishing Australia."

Translation: identical to source

GPT Explanation: identical, demonstrating full accuracy and fluency

Error: Leaving the English title untranslated constitutes a translation quality error.

Analysis: Sft-GPT misinterprets "no translation" as "perfect translation."

2.2 Bias in Understanding Linguistic Complexity

Over-tolerance on Grammar Errors

Case 10 (en-de, diff: 6.72, predicted score: 0.3208):

Source: "Grushenka inspires complete admiration and lust in both Fyodor and Dmitri Karamazov."

Translation: "Grushenka begeistert sowohl bei Fjodor als auch Dmitri Karamazow völlige Bewunderung und Lust."

GPT Explanation: missing "bei" and the use of "begeistert bei" considered a minor issue.

Error: The grammatical error in German substantially harms fluency.

Analysis: Sft-GPT lacks adequate sensitivity to the correctness in target-language.

Inadequate Handling of Mixed-language Phenomena

Case 11 (en-de, diff: 5.01, predicted score: -1.1095):

Source: "Real Time Relativity The Australian National University"

Translation: "Real Time Relativity Die Australian National University"

GPT Explanation: correctly identified as a mixed-language, ungrammatical phrase

Analysis: Sft-GPT fails to detect that most part of the sentence is not translated.

2.3 Semantic-level Evaluation Bias

Underestimation of Lexical Choice Problems

Case 12 (en-de, diff: 3.89, predicted score: 0.0003):

Source: "The common Jacobite supporters fared better than the ranking individuals"

Translation: "Den gemeinsamen Jakobiten ging es besser als den Ranking-Individuen"

GPT Explanation: detailed analysis of the issues with "gemeinsamen" and "Ranking-Individuen"

Analysis: While sft-GPT correctly identifies the problem, it still overestimates translation quality.

Tolerance of Semantic Omissions

Case 13 (en-zh, diff: 3.60, predicted score: -0.11):

Source: "During the recording of Audioslave's last album, Revelations, Morello experimented with different amplifier setups"

Translation: omits "Audioslave" and mistranslates "amplifier setups" as "amplifier sockets".

GPT Explanation: major meaning shifts though the gist remains.

Analysis: Despite acknowledging severe meaning shifts, the model assigns an overly lenient score.

C Sft-gold’s performance in Under-predicted cases

Analysis of sft-gold’s evaluation bias by MQM Error Types:

1. Accuracy Over-penalization

- **1.1 Harsh judgment on minor terminology inconsistencies**
  Case 14 (en-de, Human score: 2.21, Prediction: -0.27, Difference: -2.48):
  Source: “The sulfate minerals all contain the sulfate anion, $SO_4{}^{2-}$.”
  Translation: “Die Sulfatminerale enthalten alle Sulfatanion, $SO_4{}^{2-}$.”
  Actual quality: The German translation is basically accurate; only missing the definite article “das” before “Sulfatanion”.
  Analysis: Treated minor grammatical omission as a severe error.

- **1.2 Misjudgment of proper noun translation validity**
  Case 15 (en-de, Human score: 1.87, Prediction: -0.51, Difference: -2.38):
  Source: “Metropolitan Police and the Home Office rapidly suppressed India House”
  Translation: “Metropolitan Police und das Home Office Indien House rasch”
  Actual quality: Retaining original proper nouns is reasonable; core meaning is clear.
  Analysis: Over-penalized code-switching even in proper noun contexts.

2. Systematic Underestimation of Fluency Issues

- **2.1 Structural bias on Si-En language pair**
  Case 16 (si-en, Human score: 1.08, Prediction: -1.38, Difference: -2.46):
  Source: Complex Sinhalese political theory expression
  Translation: “Not a spoken opinion of the fascist state beginning with the ideology...”
  Actual quality: Although the word order is awkward, the core concept is conveyed clearly.
  Analysis: Over-penalized difficulty in word order transformation for Si-En pair.

- **2.2 Evaluation bias due to language distance**
  Case 17 (ne-en, Human score: 1.70, Prediction: -0.50, Difference: -2.21):
  Source: Nepali wrestling description
  Translation: “Austin was humiliated with a motion of Apartheid Custer named stunner”
  Actual quality: Contains lexical mistranslations but overall narrative structure is preserved.
  Analysis: Lacks tolerance for language pairs with large typological distance.

3. Superficial Assessment of Semantic Understanding

- **3.1 Ignoring content preservation**
  Case 18 (si-en, Human score: 1.24, Prediction: -0.88, Difference: -2.13):
  Source: Historical description of gas use during Wuhan invasion in 1938
  Translation: “from August to October 1938 during the wuchanan invasion of 375 times he was permitted to use toxic air”

Actual quality: Key historical information (time, event, frequency) is basically retained.
Analysis: Neglects information completeness, focuses excessively on lexical precision.

- **3.2 Missing understanding of functional equivalence**
Case 19 (ne-en, Human score: 2.29, Prediction: 0.25, Difference: -2.04):
Source: Travel advice about Tibetan refugee camp
Translation: "if you want to enjoy this tibetan refugee camp curling, see the workshops"
Actual quality: Core advice (visiting workshops) is clearly conveyed.
Analysis: Failed to recognize successful functional communication.

4. Misjudgment of Cultural and Contextual Adaptation

- **4.1 Ignoring specificity of religious/cultural texts**
Case 20 (si-en, Human score: 0.67, Prediction: -1.31, Difference: -1.98):
Source: Sinhalese religious tribute text
Translation: "I'll be named for you... from the distance"
Actual quality: The intent of religious text (tribute, reverence) is expressed.
Analysis: Insufficient understanding of repetitiveness and ritual features in religious texts.

- **4.2 Handling complexity of historical texts**
Case 21 (ne-en, Human score: 1.33, Prediction: -0.55, Difference: -1.88):
Source: Sanskrit names list from Five Forest Books
Translation: "Aetereae, branchian, Jupiter, stratigraphy and stratigraphy"
Actual quality: Although transliteration is inaccurate, the list structure is maintained.
Analysis: Lacks awareness of challenges in translating classical literature.

D Sft-GPT's performance in Under-predicted cases

1. Reasonable Identification of Severe Accuracy Errors

- **1.1 Accurate Judgment of Completely Distorted Semantics**
Case 22 (ne-en, Gold score: 2.28, Prediction: -2.03, Difference: -4.31):
Source: "Because they have faced the same problems."
Translation: "Because they have wrapped up the problems of the same"
GPT Explanation: Identified the severe mistranslation of "wrapped up" for "born" (suffered/experienced).
Evaluation Validity: Semantics completely distorted; -2 score is appropriate.

- **1.2 Fatal Errors in Technical Terminology**
Case 23 (si-en, Gold score: 1.75, Prediction: -2.00, Difference: -3.75):
Source: Description of System on Chip technology.
Translation: "number of documents that can be made with a single package"
GPT Explanation: Correctly identified that "functions" was mistranslated as "documents".
Evaluation Validity: Technical concept completely wrong; strict evaluation is justified.

2. Precise Identification of Grammar and Syntax Errors

- **2.1 Semantic Reversal Caused by Verb Misuse**
  Case 24 (en-de, Gold score: 1.38, Prediction: -2.30, Difference: -3.68):
  Source: “After a Federal counterattack recaptured Battery Powell”
  Translation: “Nach einem föderalen Gegenangriff rekapitulierte Van Dorn Battery Powell”
  GPT Explanation: Identified the error of “rekapitulierte” (recapitulated) vs. “recaptured”.
  Evaluation Validity: Verb misuse led to complete change of actor and nature of action.

- **2.2 Systemic Collapse of Syntactic Structure**
  Case 25 (ne-en, Gold score: 1.65, Prediction: -2.00, Difference: -3.65):
  Source: Complex description of the bill drafting process.
  Translation: “majority is prepared with the help of a methodology”
  GPT Explanation: Identified mistranslation of all key terms.
  Evaluation Validity: Entire semantic structure completely collapsed.

3. Identification of Severe Lexical Errors

- **3.1 Fatal Mistranslation of Proper Nouns and Concepts**
  Case 26 (en-zh, Gold score: 1.29, Prediction: -2.01, Difference: -3.30):
  Source: “Eisenhower asked for Graham while on his deathbed”
  Translation: “艾森豪威尔临死前向格雷厄姆求婚”
  GPT Explanation: Identified the absurd mistranslation of “asked for” as “proposed marriage”.
  Evaluation Validity: Semantics completely reversed, resulting in absurd meaning.

- **3.2 Precision Requirements in Chemical Terminology**
  Case 27 (en-zh, Gold score: 1.22, Prediction: -2.00, Difference: -3.22):
  Source: “Polonium’s isotopes tend to decay”
  Translation: “溴铵的同位素容易随着... 衰变”
  GPT Explanation: Identified the severity of completely wrong chemical element.
  Evaluation Validity: Scientific terminology errors are fatal in professional texts.

E Language Pair Specific Patterns

Si-En Issues (7/20 cases):

- Excessive penalty on structural transformations: Difficulty in Sinhalese-English word order conversion is over-penalized

- Language distance bias: Insufficient understanding of South Asian language structural differences

- Cultural and contextual neglect: Overlooked specificity of religious and cultural texts

- Underestimation of information preservation: Focused on surface fluency at expense of content completeness

Ne-En Issues (6/20 cases):

- Over-demand for lexical precision: Strict requirement on proper noun transliteration
- Fluency priority: Ignored success in overall meaning communication
- Cultural concept transfer: Lack of understanding of challenges in translating Sanskrit/Hindu concepts

En-De Comparison (4/20 cases):

- Proper noun strategy: Misjudged reasonableness of keeping English proper nouns
- Amplification of minor errors: Treated small grammatical omissions as serious problems
- Preference for language purity: Favored fully Germanized translations

Language Pair	sft-gold	sft-GPT
ro-en	1,350	1,370
ru-en	1,433	1,438
ne-en	1,365	–
et-en	705	709
en-zh	1,366	–
en-ta	–	705
si-en	–	1,425

Table 1: Exact sample sizes for each model and language pair.

Language Pair	sft-gold	sft-GPT
en-ta	705	–
en-gu	675	676
en-mr	2,652	2,652
si-en	1,425	–
en-hi	655	655
en-zh	–	1,371
ne-en	–	1,365

Table 2: Exact sample sizes for medium-performance language pairs.

Language Pair	sft-gold	sft-GPT
en-de	1,400	1,398
en-te	706	709

Table 3: Exact sample sizes for low-performance language pairs.

Bibliography

H. Akoglu. User's guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3):91–93, 2018.

J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu. Qwen technical report, 2023a. URL https://arxiv.org/abs/2309.16609.

J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023b. URL https://arxiv.org/abs/2308.12966.

H. G. Chi, F. Pesce, W. Chang, O. Rudovic, A. Argueta, S. Braun, V. Garg, and A. H. Abdelaziz. Adaptive knowledge distillation for device-directed speech detection, 2025. URL https://arxiv.org/abs/2508.02801.

Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models, 2023. URL https://arxiv.org/abs/2311.07919.

Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou. Qwen2-audio technical report, 2024. URL https://arxiv.org/abs/2407.10759.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

P. Fernandes, D. Deutsch, M. Finkelstein, P. Riley, A. F. T. Martins, G. Neubig, A. Garg, J. H. Clark, M. Freitag, and O. Firat. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation, 2023. URL https://arxiv.org/abs/2308.07286.

Y. Fu, H. Peng, L. Ou, A. Sabharwal, and T. Khot. Specializing smaller language models towards multi-step reasoning. In *International Conference on Machine Learning*, pages 10421–10430. PMLR, 2023.

A. Gajbhiye, M. Fomicheva, F. Alva-Manchego, F. Blain, A. Obamuyide, N. Aletras, and L. Specia. Knowledge distillation for quality estimation. *arXiv preprint arXiv:2107.00411*, 2021.

T. Glushkova, C. Zerva, R. Rei, and A. F. Martins. Uncertainty-aware machine translation evaluation. *arXiv preprint arXiv:2109.06352*, 2021.

N. M. Guerreiro, R. Rei, D. v. Stigt, L. Coheur, P. Colombo, and A. F. Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995, 2024.

P. He, X. Liu, J. Gao, and W. Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021. URL `https://arxiv.org/abs/2006.03654`.

G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network, 2015. URL `https://arxiv.org/abs/1503.02531`.

C.-Y. Hsieh, C.-L. Li, C.-K. Yeh, H. Nakhost, Y. Fujii, A. Ratner, R. Krishna, C.-Y. Lee, and T. Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*, 2023.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models, 2021. URL `https://arxiv.org/abs/2106.09685`.

T. Huang, S. You, F. Wang, C. Qian, and C. Xu. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35:33716–33727, 2022.

F. Kepler, J. Trénous, M. Treviso, M. Vera, and A. F. Martins. Openkiwi: An open source framework for quality estimation. *arXiv preprint arXiv:1902.08646*, 2019.

T. Kocmi and C. Federmann. Gemba-mqm: Detecting translation quality error spans with gpt-4, 2023. URL `https://arxiv.org/abs/2310.13988`.

T. Kocmi, V. Zouhar, E. Avramidis, R. Grundkiewicz, M. Karpinska, M. Popović, M. Sachan, and M. Shmatova. Error span annotation: A balanced approach for human evaluation of machine translation, 2024. URL `https://arxiv.org/abs/2406.11580`.

R. Labadie-Tamayo, D. Slijepčević, X. Chen, A. J. Böck, A. Babic, L. Freimann, and C. A. M. Zeppelzauer. Distilling knowledge from large language models: A concept bottleneck model for hate and counter speech recognition, 2025. URL `https://arxiv.org/abs/2508.08274`.

D. Larionov, M. Seleznyov, V. Viskov, A. Panchenko, and S. Eger. xcomet-lite: Bridging the gap between efficiency and quality in learned mt evaluation metrics. *arXiv preprint arXiv:2406.14553*, 2024.

S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198, 2020.

I. M. Moosa, R. Zhang, and W. Yin. Mt-ranker: Reference-free machine translation evaluation by inter-system ranking. *arXiv preprint arXiv:2401.17099*, 2024.

Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

T. Ranasinghe, C. Orasan, and R. Mitkov. Transquest at wmt2020: Sentence-level direct assessment. *arXiv preprint arXiv:2010.05318*, 2020.

R. Rei, C. Stewart, A. C. Farinha, and A. Lavie. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*, 2020.

R. Rei, A. C. Farinha, C. Zerva, D. Van Stigt, C. Stewart, P. Ramos, T. Glushkova, A. F. Martins, and A. Lavie. Are references really needed? unbabel-ist 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, 2021.

R. Rei, J. G. De Souza, D. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur, and A. F. Martins. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, 2022a.

R. Rei, A. C. Farinha, J. G. de Souza, P. G. Ramos, A. F. Martins, L. Coheur, and A. Lavie. Searching for COMETINHO: The little metric that could. In H. Moniz, L. Macken, A. Rufener, L. Barrault, M. R. Costa-jussà, C. Declercq, M. Koponen, E. Kemp, S. Pilos, M. L. Forcada, C. Scarton, J. Van den Bogaert, J. Daems, A. Tezcan, B. Vanroy, and M. Fonteyne, editors, *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 61–70, Ghent, Belgium, June 2022b. European Association for Machine Translation. URL `https://aclanthology.org/2022.eamt-1.9/`.

R. Rei, M. Treviso, N. M. Guerreiro, C. Zerva, A. C. Farinha, C. Maroti, J. G. De Souza, T. Glushkova, D. M. Alves, A. Lavie, et al. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. *arXiv preprint arXiv:2209.06243*, 2022c.

C. Samarinas and H. Zamani. Distillation and refinement of reasoning in small language models for document re-ranking. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pages 430–435, 2025.

J. Tang, R. Shivanna, Z. Zhao, D. Lin, A. Singh, E. H. Chi, and S. Jain. Understanding and improving knowledge distillation. *arXiv preprint arXiv:2002.03532*, 2020.

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan,

Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.