Universität Potsdam

Institut für Informatik und Computational Science

Institute for Computer Science and Computational Science

Masterarbeit | Master Thesis

“Self-Correction in LLM-Based Translation Systems”

Verfasserin:	Sandeep Uprety
Studiengang:	Data Science
Matrikelnummer:	804982
Erstgutachterin / First examiner:	Prof. Dr. Gerard de Melo
Zweitgutachterin / Second examiner:	Prof. Dr. Jonathan Pfaff

Ort, Datum	Berlin, 14.12.2025

Zusammenfassung

Große Sprachmodelle werden häufig für maschinelle Übersetzung eingesetzt und können flüssige Übersetzungen erzeugen, doch ihre Ausgaben enthalten weiterhin wiederkehrende Probleme wie ausgelassene Informationen, unpassende Wortwahl und Bedeutungsverschiebungen. Bisherige Arbeiten haben versucht, diese Probleme zu lösen, indem Modelle durch Prompting zur Überarbeitung ihrer eigenen Übersetzungen angeregt werden oder mehrere Generierungsschritte zur Inferenzzeit durchgeführt werden.

Diese Arbeit untersucht einen trainingsbasierten Ansatz zur Selbstkorrektur in LLM-basierten Übersetzungssystemen. Anstatt auf Prompting zur Inferenzzeit zu setzen, werden Modelle auf expliziten Beispielen ihrer eigenen Übersetzungsfehler feinabgestimmt. Die Trainingsdaten umfassen zwei Arten von Beispielen. Für fehlerhafte Übersetzungen enthält jedes Beispiel den Quellsatz, die Ausgabe des Modells, eine Analyse des Problems und die Referenzübersetzung als Korrektur. Um den Datensatz auszugleichen, werden auch korrekte Satzpaare aus dem Parallelkorpus hinzugefügt, wobei die Analyse lediglich bestätigt, dass die Übersetzung korrekt ist. Das Training erfolgt in zwei Stufen. Zunächst werden die Modelle auf parallelen Standarddaten feinabgestimmt, um grundlegendes Übersetzungsverhalten zu erlernen, und anschließend auf dem Selbstkorrektur-Datensatz mit parametereffizienten LoRA-Adaptern weiter trainiert. Die Experimente konzentrieren sich auf Deutsch-Englisch und Chinesisch-Englisch mit Modellen wie LLaMA-2, LLaMA-3, Qwen 2.5 und Mistral.

Die Selbstkorrektur-Feinabstimmung verbesserte die Übersetzungsqualität bei einigen Modellen. Qwen 2.5 verbesserte sich beispielsweise bei Chinesisch-Englisch von 32,04 auf 33,72 BLEU und von 0,842 auf 0,862 COMET. Andere Modelle zeigten geringere Verbesserungen oder verschlechterten sich teilweise. Die manuelle Evaluation zeigte, dass korrigierte Übersetzungen in der Regel besser waren, wenn das Modell einen Fehler in seiner initialen Ausgabe korrekt identifiziert hatte. Insgesamt zeigen die Ergebnisse, dass zumindest ein Teil des Selbstkorrekturverhaltens während des Trainings erlernt werden kann, anstatt auf Prompting zur Inferenzzeit angewiesen zu sein.

Schlüsselwörter: Große Sprachmodelle, Maschinelle Übersetzung, Selbstkorrektur, Feinabstimmung, LoRA, Fehleranalyse, COMET, BLEU

Abstract

Large language models are commonly used for machine translation and can generate fluent translations, yet their output still contains recurring problems such as omitting information, inappropriate lexical choices, and meaning shifts. Previous work has attempted to address these issues by prompting models to revise their own translations or by performing multiple generation steps at inference time.

This thesis evaluates a training-based approach to self-correction in LLM-based translation systems. Models are fine-tuned on examples of their own translation errors instead of relying on inference-time prompting. The training data includes two types of examples. For translations with errors, each example contains the source, the model's output, an analysis explaining the problem, and the reference as the correction. To balance the dataset, correct sentence pairs from the parallel corpus are also included where the reference translation serves as both the initial and corrected translation, with the analysis simply stating that the translation accurately captures the meaning. Training is carried out in two stages. Models are first fine-tuned on standard parallel data to learn basic translation behaviour and are then further fine-tuned on the self-correction dataset using parameter-efficient LoRA adapters. Experiments focus primarily on German-English and Chinese-English translation using models such as LLaMA-2, LLaMA-3, Qwen 2.5, and Mistral.

Self-correction fine-tuning improved translation quality for some models. Qwen 2.5 on Chinese-English, for instance, went from 32.04 to 33.72 BLEU and from 0.842 to 0.862 COMET after training. Other models improved less, and in some cases got worse. Manual evaluation showed that corrected translations were generally better when the model had correctly identified an error in its initial output. Overall, the findings indicate that at least part of self-correction behaviour can be learned during training rather than relying on prompting at inference time.

Keywords: Large Language Models, Machine Translation, Self-Correction, Fine-Tuning, LoRA, Error analysis, COMET, BLEU

ACKNOWLEDGMENT

I would like to thank Prof. Dr. Gerard de Melo and Prof. Dr. Jonathan Pfaff for their support and for giving me the opportunity to pursue this research.

I owe a special thanks to Dr. Jingyi Zhang, who generously shared her time and expertise despite her own demanding commitments. Our regular discussions were invaluable, and her feedback shaped much of this work.

I would also like to acknowledge the Hasso-Plattner-Institut for providing the GPU resources that made this research possible.

Finally, I thank my family and friends for their encouragement and patience throughout this journey.

Table of Contents

LIST OF ABBREVIATIONS

BLEU	Bilingual Evaluation Understudy
chrF	Character n-gram F-score
COMET	Crosslingual Optimized Metric for Evaluation of Translation
TER	Translation Edit Rate
DE-EN	German to English
ZH-EN	Chinese to English
GPU	Graphics Processing Unit
LLM	Large Language Model
LoRA	Low-Rank Adaptation
MQM	Multidimensional Quality Metrics
MT	Machine Translation
NLP	Natural Language Processing
NMT	Neural Machine Translation
PEFT	Parameter-Efficient Fine-Tuning
QE	Quality Estimation
SC	Self-Correction
WMT	Workshop on Machine Translation
A100	NVIDIA A100 GPU (used for training and evaluation)
ALMA	Adaptive Language Model Alignment
TasTe	Teaching Large Language Models to Translate through Self-Reflection
Europarl	European Parliament Parallel Corpus

LIST OF TABLES

1. INTRODUCTION

1.1 Background and Motivation

Large Language Models have improved natural language processing, and now perform well on many tasks, including machine translation. Models such as LLaMA, GPT-4 and their successors have achieved translation quality that equals or even exceeds traditional neural machine translation systems. These advancements have led to widespread deployment of LLM-based translation in commercial and research settings.

Despite these improvements, LLM-based translation continues to generate errors. Koehn & Knowles (2017) identified six key challenges for neural machine translation, such as domain mismatch, rare words and word alignment. More recent work has shown persistent error patterns in LLM outputs, including mistranslated words, omitted phrases, incorrect handling of named entities and grammatical errors. These errors hurt reliability, especially in legal and medical translation where precision matters.

Self-correction has emerged as one of the approaches for addressing these limitations. Wang et al. (2024) introduced TasTe, where the model first translates, then predicts the quality of its output, and uses that prediction to guide a revision. Similarly, TEaR (Feng et al., 2025) and Self-Refine (Madaan et al., 2023) use multiple forward passes or feedback loops at inference time. These approaches add computational cost during deployment.

In this thesis, we propose a different strategy for LLM self-correction where we teach the model to correct itself as part of its training. The main idea is to fine-tune the model on examples that pair flawed translations with analyses that identify the problems and show how to fix them. By exposing models to these examples during training, the hypothesis is that models can learn to avoid these errors or recognise and correct them during generation.

1.2 Problem Statement

Current approaches to self-correction in machine translation face several limitations.

Prompt-based methods require multiple inference passes: The model has to translate, then evaluate its own work and then produce a revised translation. This multiplies computational cost and latency, making it impractical for many applications.

Self-generated feedback is unreliable, as documented by Huang et al. (2024). When models generate their own feedback without external feedback, they often fail to identify actual errors or introduce new ones.

Existing fine-tuning approaches lack the detailed error information. While some work has explored fine-tuning to improve translation quality, such as ALMA-R’s use of preference pairs (Xu et al., 2024b), these methods generally do not give the model explicit information about what is wrong with a translation or why. Instead, the model must learn error patterns indirectly from examples where one translation is preferred over another.

1.3 Research Questions

The question this thesis investigates is whether fine-tuning LLMs on examples of translation errors and their analyses can teach them to self-correct. Rather than relying on prompt-based methods at inference time, the goal is to embed self-correction capabilities directly through training.

To answer this, several related questions need to be addressed. First, does self-correction training actually improve translation quality? This is measured by comparing models before and after self-correction training using both automatic metrics and manual evaluation. Second, do different model architectures respond differently to this training approach? The experiments include LLaMA-2, LLaMA-3, Qwen 2.5 and Mistral to see if some models learn self-correction better than others. Third, does the approach work across different language pairs? Testing on both German-English and Chinese-English helps determine whether the method generalises.

We also explore practical training choices. Is the two-stage approach (WMT fine-tuning followed by self-correction training) necessary, or can models learn directly from self-correction data? And what is the right balance between error examples and clean examples in the training data? These questions are addressed by comparing different training configurations.

Finally, manual evaluation checks whether the metric improvements reflect real quality gains. For Qwen 2.5, 100 random samples from each language pair were evaluated. Additional evaluations of LLaMA-2 and Mistral on German-English (100 samples each) were conducted to compare model-specific behaviour.

1.4 Approach Overview

The approach we use is a two-stage training pipeline. In the first stage, models are fine-tuned on parallel translation data from WMT. This teaches them basic translation ability. In the second stage, models are trained on a self-correction dataset that contains examples of flawed translations paired with analyses and corrections.

Each training example in the self-correction dataset has four parts: a source sentence, an initial translation (which may contain errors), an analysis of the errors and the corrected translation. The model learns to generate all three output components given the source. The analysis is manually written and explains the specific problems in the translation. Error type labels like “semantic error” or “omission” were used to organise and balance the dataset during construction, but they are not included in the training data itself.

To build the dataset, translations with COMET scores between 0.5 and 0.8 were selected. These are translations that have problems but are not completely broken. Each row was then manually annotated with an explanation of the specific issues. The dataset also includes clean examples where the initial translation is already correct, and the analysis simply mentions this. This prevents the model from learning to analyse and correct everything regardless of quality.

Experiments tested two different mixes of error and clean examples: 70% errors with 30% clean, and 50% errors with 50% clean.

All training uses LoRA for parameter-efficient fine-tuning. This allows training large models without updating all parameters, also making experiments feasible on available hardware.

1.5 Contributions

This thesis presents a methodology for teaching LLMs to self-correct through fine-tuning. Instead of just showing models correct and incorrect translation pairs, the approach trains them on explicit feedback about translation errors. This is different from preference-based methods that rely on implicit learning.

The thesis also provides results across different models and language pairs. Four different architectures were tested, covering both base and instruction-tuned models. Experiments covered both German-English and Chinese-English translation. Manual evaluation of three models on the same language pair (German-English) shows that self-correction effectiveness depends on the model architecture, not just the language pair.

Another contribution is the self-correction dataset itself. It contains source sentences, initial translations, detailed analyses and corrected outputs for both language pairs. Other researchers working on similar problems may find this dataset helpful.

Finally, the thesis provides practical insights about training configurations. The experiments show that two-stage training (WMT followed by self-correction) consistently outperforms training directly on self-correction data. They also show that the optimal ratio of error examples to clean examples is model-dependent. Some models like Qwen 2.5 performed better with 70% errors, while others worked better with a 50-50 balance.

1.6 Thesis Structure

The thesis begins by covering the technical background needed to understand the experiments. This includes the history of neural machine translation, how large language models work, the specific models used in this research, how LoRA enables efficient fine-tuning, the evaluation metrics used to measure translation quality, and a brief introduction to self-correction.

Next is a review of related work, covering previous research on self-correction in LLMs, machine translation with large language models, and fine-tuning strategies.

The methodology section then describes how the experiments were set up. It explains how the training data was prepared, how the self-correction dataset was constructed, and how the two-stage training pipeline works. It also covers the evaluation procedure for both automatic metrics and manual assessment.

The results section presents the experimental findings. This includes automatic metrics across all models and language pairs, comparisons between different training configurations and the outcomes of the manual evaluation.

A discussion follows that interprets these results. It examines why some models benefited more from self-correction training than others, considers the limitations of the approach, and reflects on what the findings mean for future work in this area.

The thesis ends with a summary of findings and recommendations for future work on self-correction training.

2. BACKGROUND

This chapter covers the background needed for the rest of the thesis. It begins with neural machine translation and how the field moved from statistical methods to neural networks and then to large language models. The chapter then describes the four models used in the experiments, explains LoRA for efficient fine-tuning, and introduces the metrics for evaluating translation quality.

2.1 Neural Machine Translation

Statistical machine translation dominated the field from the 1990s until the mid-2010s. These systems used probabilistic models to learn patterns from parallel text. Brown et al. (1993) established the mathematical foundation with the IBM Models, treating the target sentence as a noisy channel corruption of the source and splitting the problem into separate translation and language models. Phrase-based statistical machine translation (Koehn et al., 2003) improved on this by translating contiguous word sequences rather than individual words, which helped capture local reordering and idiomatic expressions.

Neural machine translation moved away from these statistical methods and used end-to-end neural networks instead. Sutskever et al. (2014) introduced the sequence-to-sequence architecture with encoder-decoder recurrent networks. The encoder reads the source sentence and compresses it into a fixed-length vector. The decoder then generates the target sentence from this vector. The problem was that all information about the source sentence had to fit into a single vector. Cho et al. (2014) analysed this limitation and showed that translation quality dropped as sentences got longer.

Bahdanau et al. (2015) addressed this with the attention mechanism, one of the most influential innovations in modern deep learning. Instead of relying on a single compressed vector, attention allows the decoder to look back at all encoder hidden states and focus on relevant parts of the source sentence at each generation step. The model learns which parts of the source sentence matter most for each word it produces. With attention, the degradation in translation quality for longer sentences was substantially reduced, and the approach became standard in neural machine translation.

Vaswani et al. (2017) proposed the Transformer architecture, which removed recurrence entirely and relied only on attention. Unlike RNNs that process words one at a time, Transformers process all positions in a sequence at once, making training much faster on GPUs. The architecture stacks multiple encoder and decoder layers. Each layer contains multi-head self-attention and position-wise feed-forward sublayers. Multi-head attention runs several attention operations in parallel with different learned projections, allowing the model to capture different types of relationships in the text at the same time. The attention function is computed as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Here, Q, K, and V are query, key, and value projections of the input embeddings. The scaling factor $\sqrt{d_k}$ prevents the dot products from getting too large, which would push softmax into regions where gradients vanish.

Another important development was subword tokenisation. Earlier systems worked with whole words, so any word not in the vocabulary was a problem. Sennrich et al. (2016) introduced Byte Pair Encoding (BPE) for machine translation to deal with this. The algorithm starts with individual characters and keeps merging the most frequent pairs until it reaches a desired vocabulary size. Words that appear often in the training data stay intact, while rare words get split into subwords. German benefits from this because compound words like "Krankenversicherung" can be broken into "Kranken" and "versicherung", both of which the model has likely seen before. This way, the model can handle words it did not encounter during training.

More recently, decoder-only large language models have become another way to do translation. Instead of training separate encoder-decoder models, these systems translate through prompting. Given a source sentence and an instruction, they generate the translation one token at a time. The models pick up linguistic knowledge from pre-training on massive multilingual text. When researchers first tried using LLMs for translation, the results were not as good as dedicated NMT systems. Xu et al. (2024a) showed that fine-tuning LLMs in two stages, first on monolingual data and then on parallel data, lets smaller models compete with much larger ones. Alves et al. (2024) developed Tower, which took a similar direction, training a model that can do translation, post-editing, and quality estimation.

2.2 Large Language Models

Large language models are neural networks trained to predict the next word in a sequence. By doing this over huge amounts of text, the models learn grammar, meaning, facts, and how to solve various tasks. Modern LLMs have billions of parameters and are trained on trillions of tokens. This scale is what allows them to handle tasks like translation, summarisation, and question answering without needing a separate architecture for each task.

The GPT series established the decoder-only approach that most LLMs use today. Radford et al. (2018) introduced GPT, showing that unsupervised pre-training followed by supervised fine-tuning works well across different tasks. GPT-2 (Radford et al., 2019) scaled this up and found that language models could do tasks without fine-tuning if given the right prompts. GPT-3 (Brown et al., 2020) went further with 175 billion parameters and showed what is called in-context learning. The model could perform tasks just from examples in the prompt, without any weight updates. This ability came from scale and made few-shot learning possible, where models adapt to new tasks from just a handful of examples.

Instruction tuning and reinforcement learning from human feedback turned out to be important for making LLMs useful in practice. Ouyang et al. (2022) introduced InstructGPT by collecting human feedback on model outputs and using it to fine-tune the model. The result was a model that people preferred to GPT-3, despite being smaller. This was an important finding because it showed that how you train a model matters just as much as its size.

Modern LLMs are built on the Transformer decoder architecture. Each layer has masked multi-head self-attention and a position-wise feed-forward network. When generating text, the model produces one token at a time using decoding strategies like beam search or nucleus sampling. Text is broken into subwords using tokenizers, and vocabulary sizes range from 32,000 to over 150,000 tokens, depending on the model.

2.3 Models Used in This Thesis

For the experiments, four models were selected: LLaMA-2, LLaMA-3, Qwen 2.5, and Mistral. All four are open-source and have between 7 and 8 billion parameters. The main reason for choosing these models is that they represent different training approaches, multilingual capabilities, and architectural innovations, and all can be fine-tuned on available hardware.

The second generation of Meta's LLaMA family, LLaMA-2 7B (Touvron et al., 2023) was introduced in July of that same year. To use less memory, the larger models (like LLaMA-2 70B) make use of Grouped Query Attention. Each head in standard multi-head attention has a unique query, key, and value matrix. Grouped Query Attention shares key and value matrices across multiple query heads, which saves memory during inference without hurting performance much. The model was trained on 2 trillion tokens with a vocabulary of 32,000 tokens and supports context lengths up to 4,096 tokens. One limitation is that the training data is mostly English, with other languages making up only about 10% of the corpus.

LLaMA-3 8B (Meta, 2024) increased the training data to over 15 trillion tokens, up from 2 trillion in LLaMA-2. They also expanded the vocabulary to 128,256 tokens. A larger vocabulary means fewer tokens are needed for the same text, which helps especially for non-English languages. Context length doubled to 8,192 tokens.

Qwen 2.5 7B (Yang et al., 2024) comes from Alibaba's Qwen Team. Unlike the LLaMA models, it was designed from the start to handle many languages. It has the largest vocabulary of the four models at 151,643 tokens, which helps with non-Latin scripts and complex morphology. The model supports over 29 languages, with particularly strong performance in Chinese. The architecture uses Rotary Position Embeddings for better handling of sequence position and SwiGLU activation in feed-forward layers. The model supports context lengths up to 128,000 tokens, the longest among the evaluated models.

Mistral 7B (Jiang et al., 2023) was released in September 2023 and introduced Sliding Window Attention. Instead of attending to all previous tokens, attention is limited to a fixed window of 4,096 tokens. This reduces memory usage. The local window might seem limiting, but stacking multiple layers allows the model to access information from further back.

2.4 Parameter-Efficient Fine-Tuning with LoRA

Fine-tuning large language models takes a lot of memory. A 7-billion parameter model in 16-bit precision needs about 14 GB just to store the weights. But training also requires gradients and optimiser states. With AdamW, the optimiser keeps two states per parameter, which roughly triples the memory needed. On top of that, activations for backpropagation add more. A full fine-tuning of a 7B model can require 80 GB or more of VRAM, which is beyond what most research hardware can handle.

Parameter-efficient fine-tuning methods deal with this by training only a small part of the model while keeping the rest frozen. There are several approaches, including adapter modules (Houlsby et al., 2019) that insert small trainable layers between the frozen blocks. Prefix tuning (Li & Liang, 2021) adds learnable vectors to the beginning of the input.

Low-Rank Adaptation or LoRA (Hu et al., 2022) is now one of the most popular methods for parameter-efficient fine-tuning. The idea is that weight updates during fine-tuning do not need the full dimensionality of the original matrix. LoRA keeps the original weights W frozen and adds a low-rank update BA:

$$W' = W + BA$$

B and A are much smaller matrices. If W is 4,096 × 4,096, that is about 16.7 million parameters. With LoRA at rank 16, you only train 4,096 × 16 + 16 × 4,096 = 131,072 parameters. This is over 99% fewer parameters.

Matrix B is initialised to zero, ensuring the model starts identically to the pre-trained version, while Matrix A uses random Gaussian initialisation. A scaling factor α controls the magnitude of the update, which helps with learning rate tuning. At inference time, LoRA weights can be merged with the original weights through simple addition, eliminating any additional latency.

In this thesis, LoRA is applied to the attention projection matrices (q_proj, k_proj, v_proj, o_proj) with rank 16 and α = 32. This makes it possible to run experiments across multiple models and configurations on a single GPU, using A100, A40, or V100, depending on availability.

2.5 Evaluation Metrics

Evaluating machine translation is difficult because there is usually more than one correct way to translate a sentence. Automatic metrics give reproducible results and scale easily, but they do not always match human judgments. This thesis uses four evaluation metrics, each measuring different aspect of translation quality. BLEU, chrF, and TER are computed using the sacreBLEU library (Post, 2018), and COMET uses the framework from Rei et al. (2020).

BLEU (Bilingual Evaluation Understudy), introduced by Papineni et al. (2002), is still the most commonly reported metric in machine translation. The metric looks at how many n-grams (for n = 1 to 4) in the translation match the reference, then combines these into a single score using a geometric mean. Short translations get penalised. Scores are typically reported on a 0 to 100 scale, where higher is better. The problem with BLEU is that it cannot recognise synonyms or valid paraphrases (Freitag et al., 2022). Because so many papers report BLEU scores, it remains helpful in comparing with prior work.

chrF (character n-gram F-score), proposed by Popović (2015), evaluates translations at the character level rather than the word level by computing an F-score over character n-grams. This makes it more robust to morphological variations and less sensitive to tokenisation. For a language like German, with many compound words, this is useful. If a model outputs "transportation" instead of "transport", BLEU treats it as a complete mismatch, but chrF sees the character overlap. chrF also correlates well with human judgments.

TER (Translation Edit Rate), introduced by Snover et al. (2006), measures the number of edit operations. Given a translation and a reference, it counts the number of insertions, deletions, substitutions, and phrase shifts needed to go from one to the other. This count is divided by the reference length. With TER, lower scores mean better translations.

COMET (Crosslingual Optimized Metric for Evaluation of Translation) was proposed by Rei et al. (2020) and takes a neural approach. It encodes the source, translation, and reference using cross-lingual embeddings and predicts a quality score through a regression head. Because it works in a continuous vector space, COMET can recognise that “car” and “automobile” mean the same thing, since their embeddings are close together. This is something surface-level metrics like BLEU cannot do. COMET correlates much better with human judgments than traditional metrics. This thesis uses the wmt22-comet-da model, which produces scores between 0 and 1.

2.6 Self-Correction in Language Models

Self-correction is when a model tries to improve its own output, either by generating feedback or by refining the output in multiple passes. Researchers have explored this because it could lead to more reliable results. Madaan et al. (2023) showed advances in tasks such as dialogue generation and code optimisation by demonstrating that a single model could carry out initial generation, feedback provision, and refinement in an iterative loop.

The effectiveness of self-correction is still an open question. Without external feedback, models often change correct answers to wrong ones or miss real mistakes (Huang et al., 2024). This suggests that prompting alone may not be enough for reliable self-correction.

This thesis takes a different approach. Instead of relying on prompts to get self-correction at inference time, the idea is to fine-tune models on explicit error analyses. The model sees examples of errors, explanations of what went wrong, and corrected outputs. The goal is to teach the model to recognise mistakes and fix them.

3. RELATED WORK

This chapter reviews prior work related to the thesis. The main goal is to provide context on LLM-based translation and self-correction before moving to the methodology. The first section examines how Large Language Models perform on translation tasks, and the next section covers self-correction in language models more generally. The third section focuses on self-correction methods built specifically for translation. The chapter concludes by identifying limitations in current methods and outlining the research gaps that this thesis aims to address.

3.1 LLM-Based Machine Translation

The application of Large Language Models to translation through prompting emerged from research into few-shot learning. Brown et al. (2020) demonstrated that GPT-3 could perform translation simply by processing a few examples in the prompt, without any fine-tuning on parallel data. The model learned to follow the pattern provided in the context and generate translations for new sentences. Although GPT-3 was not trained as a translation system, Brown et al. (2020) argue that this capability arises from large-scale language modelling, rather than from explicit supervision for translation.

Early LLM-based translation still did not match the performance of dedicated neural machine translation systems. Hendy et al. (2023) compared GPT models across 18 language pairs and found that GPT-3.5 performed competitively on high-resource pairs like German-English, but struggled with low-resource languages and particularly when translating into non-English targets. To close this gap, researchers started fine-tuning LLMs specifically for translation.

Xu et al. (2024a) developed ALMA to improve LLM translation through fine-tuning. Their idea was to first train the LLaMA model on monolingual data, and only then fine-tune it on parallel translation data. Fine-tuning directly on parallel data did not work as well. The 7B parameter ALMA model outperformed much larger systems like GPT-3.5 on translation benchmarks. They later released ALMA-R (Xu et al., 2024b), which used a novel approach called Contrastive Preference Optimization (CPO) to train the model on preference pairs of better and worse translations.

Alves et al. (2024) took a similar direction with Tower. They trained a LLaMA-2 model on multilingual data and then fine-tuned it for translation-related tasks. Beyond translation, Tower can also do post-editing and quality estimation. Both ALMA and Tower show that fine-tuning makes a real difference for LLM translation.

Wu et al. (2024) examined document-level translation using fine-tuned LLMs. They found that parameter-efficient fine-tuning can outperform Google Translate when translating into English but performs worse when translating from English to other languages. Zhu et al. (2024) found that LLMs can translate well after being fine-tuned on as few as 32 parallel sentences. This points to fine-tuning being more about teaching the format than teaching translation itself.

Fine-tuning for translation comes with trade-offs. Stap et al. (2024) found what they call a fine-tuning paradox. While fine-tuning improves general translation quality, it can hurt other abilities the model had before. Models fine-tuned for translation became worse at following instructions about formality or style. They also showed weaker document-level coherence. So improving translation through fine-tuning may come at the cost of flexibility.

Not all models see the same languages during pre-training. LLaMA-2's training data is roughly 90% English, and German makes up less than 1% (Touvron et al., 2023). Qwen 2.5 was trained on 18 trillion tokens across 29 languages (Qwen Team, 2024). How much multilingual data a model sees affects how well it handles translation.

LLM-based translation still faces persistent challenges. Koehn & Knowles (2017) documented six issues for neural translation: domain mismatch, amount of training data, rare words, long sentences, word alignment, and beam search. Hallucination is a particular concern. Guerreiro et al. (2023) found that large multilingual models hallucinate differently than traditional systems. Sometimes the model generates text in the wrong language entirely. In other cases, it adds content that does not appear in the source. Undertranslation is another problem, especially with long or complex inputs, where the model skips parts of the source sentence (Shao et al., 2024).

3.2 Self-Correction in Large Language Models

The idea behind self-correction is simple. If a model can review its own output and fix mistakes, it could produce better results without needing human feedback. This has made self-correction an appealing research direction. A model that catches its own errors would be more useful in practice and require less supervision.

But whether this works depends on how the correction happens. Some methods rely on the model alone to spot and fix problems. Others give the model access to tools or outside information that can verify whether the output is correct. The difference between these two approaches turns out to be important, and much of the recent debate in this area comes down to which type of correction researchers are testing.

3.2.1 Intrinsic versus Extrinsic Self-Correction

Kamoi et al. (2024) drew a useful distinction between two types of self-correction. Intrinsic self-correction relies entirely on the model itself. The model generates an output, reviews it and then tries to improve it using nothing but its own weights. Extrinsic self-correction brings in outside help. The model might use a calculator to check its math, run code to verify it works, or look up facts with a search engine.

The distinction matters because intrinsic self-correction asks the model to identify errors that arose from its own internal reasoning process. If the model was confident enough to produce a wrong answer, asking it to review that answer often just produces confident agreement. The same weights that generated the mistake may also struggle to recognise it as an error. Extrinsic approaches mitigate this limitation by providing an external verification signal that the model cannot reliably generate on its own. A calculator gives the right answer regardless of what the model believes.

3.2.2 Key Frameworks and Methods

Madaan et al. (2023) introduced Self-Refine, one of the most cited approaches in this space. The framework uses a single LLM to play three roles. First, the model generates an initial output. Then it writes feedback about that output, pointing out potential problems. Finally, it produces a revised version based on its own feedback. This loop can repeat until the output seems good enough. Self-Refine showed improvements on tasks like dialogue generation and code-related tasks, though the gains on reasoning tasks were smaller.

Shinn et al. (2023) took a different direction with Reflexion. Instead of revising outputs immediately, Reflexion has the model reflect on failed attempts and store those reflections in memory. When the model tries the task again, it can read its past reflections and avoid repeating the same mistakes. This worked well for tasks where success and failure are easy to define, such as coding problems where you can run tests to see whether the code works.

Chain-of-thought prompting (Wei et al., 2022) is not exactly self-correction but shares some of the same goals. By asking models to show their reasoning step by step before giving a final answer, the chain-of-thought improves performance on complex problems. The model is not correcting a previous output, but it is doing something similar by working through its thinking explicitly rather than jumping straight to an answer.

3.2.3 Does Self-Correction Actually Work?

Recent work shows that the effectiveness of self-correction depends on several factors. Intrinsic self-correction, where the model reviews its own output without external feedback, has shown limited reliability. Huang et al. (2024) found that when models were asked to review and revise their answers to reasoning tasks, they often changed correct answers into incorrect ones and defended incorrect responses. This behaviour suggests that, without an independent verification signal, models struggle to reliably judge the correctness of their own outputs.

Gou et al. (2024) reported positive results by allowing language models to use external tools. In their CRITIC framework, a model can use tools such as a calculator to check mathematical computations or a search engine to verify factual claims. Instead of relying only on its own output, the model compares its answers with information from these external sources. This approach represents extrinsic self-correction, where improvements come from feedback that is provided outside the model.

Another approach is training-based self-correction. Kumar et al. (2025) showed that models trained on correction examples from other systems can fix those systems' errors but struggle to correct their own, since different models make different kinds of mistakes. To address this, they trained models on their own errors using multi-turn reinforcement learning, which led to improved self-correction performance, particularly on mathematical tasks.

Pan et al. (2024) reviewed prior work on self-correction and reached a similar conclusion. Across a wide range of tasks, pure intrinsic self-correction was found to be unreliable. Additional mechanisms are required, such as access to external tools, fine-tuning on correction data, or prompting strategies designed for specific tasks.

3.3 Self-Correction in Machine Translation

Although translation has its own methods and challenges, the main ideas about self-correction also apply to translation. Researchers have tried both training-based methods, where the model learns to correct itself while it is being improved, and inference-time methods, where the model makes its translation better by going over it several times. Each method has its own balance between how well it works and how much computer power it needs.

3.3.1 Inference-Time Approaches

A number of research teams have explored inference-time self-correction for machine translation, where a model first generates a translation and then evaluates and revises it during decoding.

The TEaR framework from Feng et al. (2025) divides inference-time self-correction into three steps. The model first generates a translation, then estimates its quality using prompts based on the MQM error taxonomy, and finally refines the translation based on the identified issues. While results across multiple language pairs were promising, the approach can require up to three forward passes (Translate-Estimate-Refine), increasing inference cost compared to single-pass translation.

Wang et al. (2024) took a more straightforward approach with TasTe. Instead of detailed error analysis, their system uses coarse quality labels to decide when a translation needs revision. This cuts down on computational cost but still leads to measurable gains. The model learns to both translate and judge its translations during training.

Automatic post-editing with large language models has also shown promise. Raunak et al. (2023) used GPT-4 with chain-of-thought prompting to revise translations produced by other systems. While this approach improved translation quality, it relies on an expensive model and requires multiple inference steps.

Chen et al. (2024) showed that iterative refinement does not consistently improve translation quality. While refinement can address certain issues in the output, additional iterations do not guarantee further gains and can sometimes degrade performance.

Berger et al. (2024) examined the effect of providing explicit error annotations before asking models to correct their translations. When errors were clearly marked, correction quality improved. In contrast, without such annotations, models often judged their own translations as correct and made few changes. This result aligns with broader findings in the self-correction literature, which show that models typically require external signals to reliably identify their mistakes.

3.3.2 Training-Based Approaches

Instead of correcting at inference time, another option is to build correction ability into the model during training. The model learns what errors look like and how to fix them, so it can do better on the first pass or at least know how to revise more effectively.

Preference-based training is one way to encourage better translation behaviour during training. ALMA-R, discussed in Section 3.1, uses Contrastive Preference Optimization to train models on pairs of translations where one output is preferred over another. While this approach teaches the model to favour higher-quality translations, the training signal remains implicit: the model learns which translation is better, but not what specific errors led to the preference. As a result, it must infer error patterns indirectly from the data.

Welleck et al. (2022) explored a more direct approach called self-corrective learning. They generated outputs from the model, identified which ones had errors, and then trained on sequences that included both the error and its correction. Models trained this way outperformed those relying on prompting alone for correction in their evaluated generation tasks.

MT-Ladder from Feng et al. (2024) builds training data from triplets. Each example has a source sentence, a flawed translation generated by an LLM, and a reference translation. The model learns to refine the flawed version toward the reference. They use a curriculum that starts with easier corrections and moves to harder ones. After training, the model can often improve translations without requiring repeated inference steps, reducing the need for iterative correction at inference time.

3.3.3 Quality Estimation and Error Analysis

Understanding what makes a translation wrong is basic to any correction effort. The MQM framework from Lommel et al. (2014) provides a taxonomy for this. Errors fall into categories like accuracy, fluency, terminology, and style, each with its own subcategories. MQM has become standard for detailed human evaluation in translation research.

Large language models can generate error analyses when prompted appropriately. Fernandes et al. (2023) found that model-generated feedback often aligned with judgments made by human annotators. Although the quality of this feedback depends on prompt design, the results suggest that LLM-generated feedback could serve as training data for correction systems.

Quality estimation is a related but separate problem. QE predicts translation quality without needing a reference translation. Fomicheva et al. (2020) built unsupervised QE methods using model uncertainty, and Kocmi & Federmann (2023) showed that LLMs can do quality estimation directly through their GEMBA framework.

Guerreiro et al. (2024) extended quality estimation for correction with xCOMET, which provides both a sentence-level quality score and token-level information indicating where errors occur. By highlighting specific words or phrases associated with errors, xCOMET offers more actionable feedback than sentence-level quality scores alone.

He et al. (2024) explored using quality estimation scores as rewards during training. While this approach led to some improvements, they found that models could exploit weaknesses in the QE system rather than genuinely improving translation quality. As a result, QE scores increased without corresponding gains in actual translation quality.

3.4 Research Gaps

The work reviewed so far points to a few things that have not been fully addressed.

Inference-time approaches like TEaR (Feng et al., 2025) and TasTe (Wang et al., 2024) require multiple passes through the model. Each additional pass increases latency and computational cost. For real-world applications where efficiency is important, a single-pass approach that embeds correction behaviour directly into the model weights is therefore likely to be more practical.

Preference-based methods like ALMA-R (Xu et al., 2024b) teach models to prefer better translations, but the signal is indirect. The model never learns what the actual problem was. Training with explicit error information could give models a clearer picture of what to look for and how to correct it.

Most correction datasets rely on synthetic errors or errors generated by other systems, which may not reflect the mistakes made by a specific model. Kumar et al. (2025) showed that this mismatch matters for self-correction, as models trained on other systems' errors struggled to correct their own. Training on errors produced by the model itself is therefore likely to provide a more targeted correction signal.

There is also limited comparison across architectures. Most papers test one or two models and report results. It is difficult to know from the literature whether self-correction training benefits different model architectures to the same extent.

3.5 Position of This Thesis

This thesis takes a training-based approach that addresses the gaps outlined above. Instead of prompting for correction at inference time, models are fine-tuned on data that contains explicit error analyses paired with corrections. Each training example includes the source sentence, the model's own flawed translation, a detailed explanation of what went wrong, and the corrected version.

This approach combines ideas from several strands of prior work. It follows a two-stage training pipeline, similar to ALMA and Tower. In the first stage, the model is trained on parallel data to establish translation ability. In the second stage, it is trained on self-correction data. Unlike preference-based methods, where the model must infer why one translation is better than another, the training data here provides explicit error information. In contrast to inference-time approaches, which require multiple passes through the model, correction behaviour is embedded directly into the model weights, enabling improved translations in a single pass.

The training data is derived from errors produced by the model itself rather than from synthetic sources. After the first training stage, the model translates a set of sentences, and translations with quality issues are identified using COMET scores. These outputs are then manually analysed to determine error types and produce corrected translations. As a result, the models are trained to learn from actual mistakes.

The experiments cover four model architectures and two language pairs. The models are LLaMA-2 7B, LLaMA-3 8B, Qwen 2.5 7B, and Mistral 7B. The language pairs are German to English and Chinese to English. German-English represents a high-resource pair where the languages share some vocabulary and similar sentence structure. Chinese-English is more distant, with a different writing system and grammar. Testing across multiple models and language pairs makes it possible to see whether self-correction training generalises or only works in specific conditions.

3.6 Summary

This chapter reviewed self-correction methods in LLMs and their application to machine translation. The main point is that intrinsic self-correction on its own has been shown to be unreliable in many settings. Models need external signals, tools, or dedicated training to reliably correct their mistakes. Inference-time approaches can improve translations but add computational cost through multiple passes. Training-based approaches embed correction ability into the model weights, which is more efficient at inference time but requires appropriate training data. This thesis addresses gaps in the current literature by training models on explicit error analyses of their own mistakes. The next chapter describes the methodology used in this thesis.

4. METHODOLOGY

This chapter describes the methodology used to investigate whether fine-tuning on error analyses can improve self-correction capabilities in LLM-based machine translation. The methodology consists of three main components: data preparation, training and evaluation.

4.1 Data Preparation

Two types of datasets were prepared: parallel translation data for WMT fine-tuning and self-correction data for learning error correction. Both were filtered for quality and checked for contamination between training and test sets.

4.1.1 Data Filtering Approach

Different filtering approaches were necessary for each language pair due to linguistic differences between the source languages.

For German-English, word-based filtering (10-40 words per sentence) was used since German uses spaces between words, making word count a straightforward measure of sentence length.

For Chinese-English, character-based filtering (50-250 characters) was used instead. Chinese does not use spaces between words, so word-based filtering would not work. Tokenising by spaces would identify most sentences as having only 1-2 words.

4.1.2 German-English Parallel Data

Data Sources: The German-English parallel data comes from two sources: Europarl v9 (European Parliament proceedings) and WMT Newstest 2008-2019 (news domain test sets from 12 years of WMT shared tasks). This combination provides domain diversity between formal political discourse and news text.

Filtering Criteria: All sentence pairs were filtered to contain 10-40 words in both source and target, and exact duplicates were removed.

The final training dataset contains 13,000 sentence pairs from Europarl and 13,000 from newstest, totalling 26,000 pairs.

Attribute	**Details**	**Size**
Training Corpora	Europarl v9 + newstest 2008-2019	26,000 pairs
Filtering Criteria	10-40 words per sentence	
Validation Set	Held out newstest	2,000 pairs
Test Set	Held out newstest	5,000 pairs

Table 1. German-English Dataset Statistics

4.1.3 Chinese-English Parallel Data

Data Sources: The Chinese-English parallel data comes from two sources: News Commentary v15 and the UN Parallel Corpus, with 15000 pairs sampled from each.

Filtering Criteria: All sentence pairs were filtered to contain 50-250 characters in both source and target, and exact duplicates were removed.

The final training dataset contains 30000 pairs.

Attribute	**Details**	**Size**
Training Corpora	News Commentary v15 + UN Parallel	30,000 pairs
Filtering Criteria	50-250 characters per sentence	
Validation Set	News Commentary + UN held-out	2,000 pairs
Test Set	News Commentary v15 held-out	5,000 pairs

Table 2. Chinese-English Dataset Statistics

4.1.4 Self-Correction Dataset Creation

The self-correction dataset was created using the WMT-fine-tuned LLaMA-2 model but the dataset itself works for all models, and every model was trained on the same data. This model was used to translate a held-out portion of both the German-English and Chinese-English corpora. The resulting translations reflect the kinds of errors the model still makes after standard fine-tuning, which makes them suitable material for building a correction-oriented dataset.

Not all translations were equally useful. To focus on cases where the model produced understandable but imperfect output, translations were filtered by COMET score. Only sentences that ranged between 0.5 and 0.8 were kept. Outputs in this range usually convey the general meaning but contain clear issues such as missing details, wrong word choices, grammatical mistakes or distortions of the source. Scores below 0.5 often indicate translations too broken to correct meaningfully, while scores above 0.8 rarely need any changes. These score ranges also matched what we saw in our initial checks, where mid-range COMET scores typically meant the translation was understandable but still needed fixing.

Each selected example was manually inspected. An error type was assigned based on the taxonomy in Section 4.1.5, and an analysis was written describing what went wrong. For the corrected translation, the reference translation from the original parallel corpus was used. Therefore, the corrected translation is always derived from the human reference and is not something the model generated.

Early analyses used a long paragraph format with phrases like “The model translation is...” and “The correct translation should be...”. This format proved ineffective, so the format was changed to a structured format with clear headers:

- Initial translation: the model output
- Analysis: a short explanation of the error
- Corrected translation: the reference from the parallel corpus

The |||END||| token used in WMT training was also removed from self-correction data, as it caused the model to stop generating too early in the multi-part format and generated broken outputs.

The dataset also includes examples where the translation was already correct. For these, the analysis is just a simple “the translation accurately captures the meaning”. This teaches the model when not to make changes.

The German-English and Chinese-English datasets differ in one area of the analysis format. For German-English, error analyses referenced the German source text directly, for example: The German “ausgetauscht” should be “exchanged” not “replaced”. After evaluating outputs from the German-English models, the Chinese-English dataset was created with analyses written entirely in English, without Chinese characters. This was done to test whether removing source language from the analysis would reduce confusion during generation.

The final data uses a prompt-completion format for fine-tuning. The prompt contains only the source sentence:

Translate the following [Source Language] to English:

{source_sentence}

The completion contains the initial translation, error analysis, and corrected translation:

Initial translation: {initial_translation}

Analysis: {error_analysis}

Corrected translation: {corrected_translation}

During training, the prompt and completion are concatenated together. The prompt tokens are masked so the model only learns to generate the completion given the prompt.

Two versions of the dataset were prepared: a balanced 50-50 split and a 70-30 version where error cases form the majority. Both use the same structure but differ in sample ratio.

4.1.5 Error Type Taxonomy

To ensure balanced representation, each error sample was classified into one of seven categories:

Error Type	**Description**

Lexical Error	Wrong word choice within the correct meaning domain (e.g., “calls” instead of “routes”)
Semantic Error	Complete meaning shift, translation conveys a fundamentally different concept than the source (e.g., “Transportation” translated as “Telecommunications”)
Omission/Addition	Content missing from the source (omission) or content added that was not in the source (addition)
Named Entity Error	Incorrect handling of names, places, organisations (e.g., “Committee” instead of “Commission”)
Grammatical Error	Incorrect grammar, syntax or sentence structure in the target language (e.g., wrong verb tenses)
Fluency Error	Unnatural or awkward phrasing, grammatically correct, but does not sound fluent.
Incomplete Translation	Translation cuts off mid-sentence

Table 3. Error Type Taxonomy

Appendix A provides examples of each error type from the German-English dataset.

4.1.6 Annotation Process

Translation errors often fall into more than one category at the same time. For example, a single error might involve both a wrong word choice and an omission. To handle this, each error was first tagged with one or more error types. Then, the category that best represented the main issue was selected as the primary classification. This selection also considered the overall distribution to keep the dataset balanced across error types.

The error-type classification was used only for dataset creation to ensure that no single error type dominated the training data. The error type labels were not included in the training data.

4.1.7 Error Distribution by Language Pair

Error Type	**Count**	**%**
Lexical Error	134	18.2%
Semantic Error	126	17.1%
Omission/Addition	117	15.9%
Named Entity Error	109	14.8%
Grammatical Error	105	14.3%

Fluency Error	101	13.7%
Incomplete Translation	44	6.0%
Total	**736**	**100%**

Table 4. German-English Error Distribution

Error Type	**Count**	**%**
Named Entity Error	142	17.8%
Semantic Error	136	17.0%
Fluency Error	131	16.4%
Lexical Error	129	16.1%
Omission/Addition	115	14.4%
Grammatical Error	88	11.0%
Incomplete Translation	59	7.4%
Total	**800**	**100%**

Table 5. Chinese-English Error Distribution

4.1.8 Self-Correction Training and Validation Split

The error samples were combined with clean translation samples and split into training and validation sets. We tested two different compositions to see which ratio works better.

The first version used a 50-50 split with equal numbers of error corrections and clean translations. This tests whether balanced exposure to both types of examples is optimal.

The second version used a 70-30 split with more error corrections than clean translations. This was created by keeping all error samples and downsampling the clean samples. The idea was to test whether more exposure to error correction examples helps the model learn better.

Dataset	**Size**	**Details**
Error Samples	800	Manual error analysis
Clean Samples	800	Correct translations
50-50 Training	1,400	700 errors + 700 clean
70-30 Training	1,000	700 errors + 300 clean
Validation Set	200	100 errors + 100 clean
Test Set	5,000	Evaluation

Table 6. Chinese-English Self-Correction Dataset

Dataset	Size	Details
Error Samples	736	Manual error analysis
Clean Samples	736	Correct translations
50-50 Training	1,288	644 errors + 644 clean
70-30 Training	920	644 errors + 276 clean
Validation Set	184	92 errors + 92 clean
Test Set	5,000	Evaluation

Table 7. German-English Self-Correction Dataset

4.1.9 Data Quality Assurance

To maintain data quality and avoid contamination:

- Train and test separation: All training, validation, and test sets were verified to have zero overlap.
- Deduplication: Duplicate sentence pairs were removed from all datasets.
- Contamination checks: Checked and confirmed no overlap between training data and evaluation data.
- Clean sample isolation: Clean samples for self-correction training are taken from separate WMT data pools with verified zero overlap with all other datasets.

4.2 Training Pipeline

The training process has two stages. First, the model is fine-tuned on parallel translation data (WMT fine-tuning) to learn basic translation. Then, it is fine-tuned again on the self-correction data to learn error correction. Both stages use LoRA (Low-Rank Adaptation) for parameter-efficient training (Hu et al., 2022).

4.2.1 Models

Four models were used to test whether the approach works across different architectures:

Model	HuggingFace ID	Parameters	Type
LLaMA-2	meta-llama/Llama-2-7b-hf	7B	Base
LLaMA-3	meta-llama/Meta-Llama-3-8B	8B	Base
Qwen 2.5	Qwen/Qwen2.5-7B-Instruct	7B	Instruct
Mistral	mistralai/Mistral-7B-Instruct-v0.1	7B	Instruct

Table 8. Models Used in Experiments

The selection includes both base models (LLaMA-2, LLaMA-3) and instruction-tuned models (Qwen 2.5, Mistral). This allows us to test whether instruction-tuned models learn self-correction differently than base models.

4.2.2 LoRA Configuration

Parameter	**Value**
Rank (r)	16
Alpha	32
Dropout	0.05
Target Modules	q_proj, k_proj, v_proj, o_proj

Table 9. LoRA Configuration (Both Stages)

4.2.3 Training Hyperparameters

Parameter	**WMT Training**	**Self-Correction**
Batch Size	2	2
Gradient Accumulation Steps	8	8
Effective Batch Size	16	16
Learning Rate	1e-4	1e-4
Number of Epochs	3	3
Max Sequence Length	512 tokens	768 tokens
Optimizer	paged_adamw_32bit	paged_adamw_32bit
LR Scheduler	cosine	cosine
Warmup Ratio	0.05	0.05
Weight Decay	0.01	0.01
Precision	bfloat16	bfloat16

Table 10. Training Hyperparameters

Note: Self-correction uses a longer sequence length (768 vs 512) because the output includes the initial translation, analysis, and corrected translation

4.2.4 Stage 1: WMT Fine-tuning

The first stage trains the base model to translate from the source language to English using the parallel data described in Section 4.1.

The prompt format is:

Task: Translate the [Source Language] sentence into English.

[Source Language]: {source_sentence}

English:

The completion format is:

{target_sentence}|||END|||

Here [Source Language] is replaced with "German" or "Chinese" depending on the language pair.

During training, the prompt and completion are concatenated into a single sequence. The labels for the prompt tokens are set to -100, which tells the model to ignore them when computing the loss. This way, the model only learns to generate the completion.

After training, the LoRA adapter is merged with the base model weights using the PEFT library. This creates a standalone WMT-tuned model which we use as the starting point for the next stage.

4.2.5 Stage 2: Self-Correction Fine-tuning

The second stage trains the WMT-tuned model to analyse translations and correct errors when needed. Training starts from the merged WMT model, not the original base model. This way, the model already knows how to translate before learning to correct.

The prompt format is:

Translate the following [Source Language] to English:

{source_sentence}

The completion format is:

Initial translation: {initial_translation}

Analysis: {error_analysis}

Corrected translation: {corrected_translation}

Here [Source Language] is replaced with "German" or "Chinese" depending on the language pair. The model is trained to generate full completion given the prompt.

Training follows the same approach as Stage 1, with prompt tokens masked so the model only learns to generate the completion.

There are two main differences from the first stage. First, the |||END||| token is not used. Early experiments showed that including this token caused the model to stop generating too early, often cutting off the output after the initial translation or analysis. Second, the maximum sequence length is increased from 512 to 768 tokens. This is needed because the output now includes three parts (initial translation, analysis and corrected translation) instead of just the translation.

After training, the LoRA adapter is merged with the WMT model weights using the PEFT library. This self-correction model is then used directly for translation and evaluation.

4.2.6 Alternative: Direct Self-Correction Training

To test whether the two-stage approach is necessary, an alternative experiment was also conducted. In this setup, the base model is trained directly on self-correction data without the WMT fine-tuning stage. This tests whether a model can learn self-correction without first learning basic translation. The same LoRA configuration and hyperparameters are used but training starts from the original base model instead of the WMT-merged model.

Comparing the two approaches (two-stage vs direct) shows whether the WMT stage provides any benefit for learning self-correction.

4.3 Evaluation Methodology

The evaluation aims to answer two questions: does self-correction training improve translation quality, and does the model actually learn to correct its own errors?

4.3.1 Evaluation Metrics

Translation quality was evaluated using four metrics. BLEU (Papineni et al., 2002) measures n-gram overlap between a candidate translation and a reference and was computed using sacreBLEU. chrF (Popović, 2015) operates at the character level, making it more robust to morphological variation than word-based metrics. TER (Snover et al., 2006) measures the minimum number of edit operations required to transform a system output into the reference, with lower scores indicating better quality. COMET (Rei et al., 2020) is a neural evaluation metric trained on human quality judgments and has been shown to correlate more strongly with human evaluation than traditional metrics. The Unbabel/wmt22-comet-da model was used to compute COMET scores.

4.3.2 Evaluation Pipeline

Each model was evaluated at three stages. First, the original pre-trained model without any fine-tuning. Second, after WMT fine-tuning. And third, after both WMT and self-correction fine-tuning. Comparing these three stages shows how much improvement comes from translation training versus self-correction training. For generations, beam search with 4 beams was used without sampling, ensuring deterministic results. The maximum output length was set to 256 tokens for base and WMT models, and 400 tokens for self-correction models, since they have a longer output format.

Since self-correction models produce a longer output that includes the initial translation, analysis, and corrected translation, the corrected translation needs to be extracted for evaluation. This was done by searching for the “Corrected translation:” marker. If the extraction failed, the initial translation was used instead.

Each model was evaluated on 5,000 test sentences per language pair. The same test set was used for all three stages to ensure the results are directly comparable.

4.3.3 Manual Evaluation

Automated metrics do not always reflect actual translation quality, so 200 samples (100 per language pair) were randomly selected from Qwen 2.5, the best-performing model, and manually evaluated. To understand why some models showed metric degradation, LLaMA-2 and Mistral were also evaluated on German-English (100 samples each).

Each sample was categorized based on two factors. First, whether the initial translation was correct or contained errors. Second, whether the model analysed the translation or left it unchanged. For samples that were analysed, the evaluation checked whether the correction improved the translation, made it worse, or kept it the same.

Specifically, manual evaluation was done to look for:

- Whether the model correctly identified good translations and left them unchanged
- Whether the model successfully fixed errors in problematic translations
- Whether the model introduced new errors when correcting
- Whether the model missed errors it should have caught

4.4 Experimental Setup

All experiments were run on a single NVIDIA A100 GPU with 40GB memory. Training used Hugging Face Transformers with PEFT for LoRA, and bfloat16 precision. Evaluation used sacrebleu (BLEU, chrF, TER) and the Unbabel COMET implementation.

Four models were tested across two language pairs at three stages (Base, WMT and Self-Correction), plus two data compositions (50-50 and 70-30) and a direct training variant. Each configuration was evaluated on 5,000 test sentences per language pair.

4.5 Summary

This chapter described the methodology for investigating self-correction in LLM-based machine translation. The main components are:

- A data preparation pipeline that creates self-correction training data with error analyses categorised into seven error types
- Language-appropriate filtering (character-based for Chinese, word-based for German)
- A two-stage training approach: WMT fine-tuning followed by self-correction fine-tuning
- Two data compositions (50-50 balanced and 70-30 error-weighted)
- Evaluation using automated metrics and manual evaluation

The next chapter contains the results of these experiments.

5. RESULTS

This chapter presents the results from the experiments described in Chapter 4. The main question is whether self-correction training improves translation quality and whether models learn to fix their own mistakes.

The chapter begins with automatic evaluation metrics across all models and language pairs, followed by manual evaluation to check whether the metric changes reflect real improvements. It also compares different dataset configurations and examines whether the two-stage training approach outperforms direct training. In all tables throughout this chapter, bold values indicate better performance.

5.1 Automatic Evaluation Results

The tables below show results for each model at three stages: base, WMT fine-tuned and self-correction fine-tuned. For each model, the best-performing configuration (either 70-30 or 50-50 split) is shown. The full comparison of different configurations is in Section 5.3.

All four metrics are reported in the tables, but the discussion focuses primarily on BLEU and COMET. BLEU measures word-level overlap with the reference, while COMET is a neural metric trained on human judgments. Together, they capture both surface-level and semantic quality. The chrF and TER scores are included for completeness and generally follow similar patterns.

Overall, self-correction training improved results for most models on Chinese-English, with Qwen 2.5 showing the largest gains. German-English results were more mixed, with some models improving and others degrading.

5.1.1 Chinese-English Results

Table 11 shows the results for Chinese-English translation across the three training stages.

Model	Stage	BLEU	COMET	chrF	TER	Config
Qwen 2.5 (Instruct)	Base	24.89	0.823	55.07	69.96	
	WMT	32.04	0.842	57.76	56.63	
	Self-Correction	**33.72**	**0.862**	**60.43**	57.47	70-30
Mistral (Instruct)	Base	17.02	0.819	50.50	81.76	
	WMT	30.54	0.853	57.42	59.02	
	Self-Correction	**30.80**	**0.854**	**58.05**	60.72	50-50
LLaMA-2 (Base)	Base	19.74	0.829	52.58	77.79	
	WMT	31.16	0.856	57.97	58.41	
	Self-Correction	**31.27**	**0.857**	**58.38**	59.49	50-50
LLaMA-3 (Base)	Base	23.98	0.848	56.80	71.44	
	WMT	32.90	0.847	58.65	56.92	

	Self-Correction	**33.67**	**0.862**	**60.39**	57.63	50-50

Table 11. Chinese-English Translation Results (Best Configuration Per Model)

All four models improved after self-correction training, though the size of improvement varied considerably.

Qwen 2.5 showed the largest gains. Starting from a base BLEU of 24.89, WMT fine-tuning brought it to 32.04, and self-correction training pushed it further to 33.72. COMET followed a similar pattern, going from 0.823 to 0.842 after WMT and then to 0.862 after self-correction. The chrF score improved from 55.07 to 60.43, and TER dropped from 69.96 to 57.47 (lower is better). The best results came from the 70-30 configuration, though the 50-50 split performed similarly.

LLaMA-3 also improved consistently. BLEU went from 23.98 (base) to 32.90 (WMT) to 33.67 (self-correction), and COMET from 0.848 to 0.847 to 0.862. Interestingly, WMT fine-tuning slightly decreased COMET for this model, but self-correction training recovered and exceeded the original score. The 50-50 configuration worked best for LLaMA-3.

Mistral and LLaMA-2 showed smaller improvements from self-correction. For Mistral, BLEU increased from 30.54 to 30.80 and COMET from 0.853 to 0.854. For LLaMA-2, BLEU went from 31.16 to 31.27 and COMET from 0.856 to 0.857. These gains are modest, but neither model's performance decreased. Both performed better with the 50-50 split.

All models benefited more from WMT fine-tuning than from self-correction training. Mistral, for example, jumped from 17.02 to 30.54 BLEU during WMT training but only gained 0.26 more from self-correction. This is expected since WMT fine-tuning teaches general translation ability, and self-correction only adjusts it further. One exception is TER, which increased slightly for all models after self-correction. TER measures edit distance from the reference, so higher TER means the corrected translations use different wording than the reference even when semantic quality improves.

5.1.2 German-English Results

Table 12 shows the German-English results. Unlike Chinese-English, the outcomes here were mixed. Some models improved, while others got worse.

Model	**Stage**	**BLEU**	**COMET**	**chrF**	**TER**	**Config**
Qwen 2.5 (Instruct)	Base	30.63	0.831	58.62	61.82	
	WMT	32.98	0.828	56.38	53.98	
	Self-Correction	**33.76**	**0.841**	**58.64**	55.85	70-30
Mistral (Instruct)	Base	31.28	0.841	57.91	57.33	
	WMT	36.61	0.857	50.92	60.82	
	Self-Correction	33.37	0.843	58.29	55.48	70-30
LLaMA-2 (Base)	Base	32.91	0.848	59.65	55.74	

	WMT	37.52	0.860	61.48	49.82	
	Self-Correction	33.37	0.846	59.07	58.23	70-30
LLaMA-3 (Base)	Base	34.78	0.854	61.22	53.67	
	WMT	34.58	0.832	57.68	52.42	
	Self-Correction	34.08	0.846	58.91	54.15	70-30

Table 12. German-English Translation Results (Best Configuration Per Model)

Qwen 2.5 was again the strongest performer. BLEU increased from 30.63 (base) to 32.98 (WMT) to 33.76 (self-correction). Whereas for COMET, it dropped slightly during WMT training (0.831 to 0.828) but then recovered and improved after self-correction (0.841). It seems self-correction training helped restore semantic quality that WMT training had lost. The 70-30 configuration worked best for Qwen on German-English. The chrF score showed the same pattern, dropping from 58.62 to 56.38 during WMT training and recovering to 58.64 after self-correction.

LLaMA-3 showed a split result. COMET improved from 0.832 to 0.846, but BLEU dropped slightly from 34.58 to 34.08. COMET is trained on human judgments and tends to correlate better with human evaluation than BLEU (Rei et al., 2020), so the improvement in COMET despite the BLEU drop suggests the translations are not actually worse. The 70-30 configuration worked best here as well.

Mistral and LLaMA-2 both degraded after self-correction training. Mistral dropped from 36.61 to 33.37 BLEU, and LLaMA-2 from 37.52 to 33.37. Both drops are quite large. Both models had achieved strong performance after WMT fine-tuning, among the highest BLEU scores in the experiment, but self-correction training undid some of that progress. The 70-30 configuration performed better than 50-50 for both as well, but even then, the results were worse than the WMT baseline.

5.1.3 Summary of Self-Correction Improvements

Table 13 summarises the change from WMT to self-correction training for each model. Qwen improved on both language pairs, LLaMA-3 improved on COMET but not always BLEU, and Mistral and LLaMA-2 degraded on German-English.

Model	**Language Pair**	**BLEU**	**COMET**	**chrF**	**TER**
Qwen 2.5	Chinese-English	**+1.68**	**+0.020**	**+2.67**	+0.84
	German-English	**+0.78**	**+0.013**	**+2.26**	+1.87
LLaMA-3	Chinese-English	**+0.77**	**+0.015**	**+1.74**	+0.71
	German-English	-0.50	**+0.014**	**+1.23**	+1.73
Mistral	Chinese-English	+0.26	+0.001	+0.63	+1.70
	German-English	-3.24	-0.014	**+7.37**	**-5.34**
LLaMA-2	Chinese-English	+0.11	+0.001	+0.41	+1.08
	German-English	-4.15	-0.014	-2.41	+8.41

Table 13. Change in Metrics from WMT to Self-Correction Training

5.1.4 Initial vs Corrected Translations

The previous sections compared models at different training stages. This section looks at something different: within a single inference, does the self-correction model actually improve its own output?

At inference time, the self-correction model generates three parts: an initial translation, an analysis, and a corrected translation. Table 14a compares the initial and corrected translations from Qwen 2.5 on the test set.

Language Pair	BLEU	COMET	chrF	TER	Stage
Chinese-English	33.80	0.864	60.75	57.24	Initial
	33.72	0.862	60.43	57.47	Corrected
	-0.08	-0.002	-0.32	+0.23	Change
German-English	34.18	0.853	60.08	53.61	Initial
	33.76	0.841	58.64	55.85	Corrected
	-0.42	-0.012	-1.44	+2.24	Change

Table 14a. Initial vs Corrected Translation Comparison (Qwen 2.5)

The metrics suggest that corrections made things slightly worse. All scores decreased marginally (for TER, higher means worse, so the increase is also a decline).

However, the drop does not reflect a real decrease in quality. Reference-based metrics penalise any difference from the reference, even when the change is an improvement. If the model paraphrases a sentence or restructures it to be clearer, the metrics will drop because the words no longer match the reference exactly.

Manual evaluation of 100 random samples per language pair showed a different picture. For German-English, the model corrected 14 out of 15 identified errors without making any of them worse. For Chinese-English, 6 translations improved while only 2 degraded, a net gain of 4. These findings suggest that the model is making genuine improvements that automatic metrics cannot capture.

The comparison above shows initial versus corrected translations from the same self-correction model. However, the self-correction model's initial translation is not identical to the WMT model's output. Self-correction training changes how the model translates, even before any correction is applied. Table 14b compares these three stages for German-English.

Model	Language	WMT BLEU	SC Initial BLEU	SC Final BLEU
Qwen 2.5	ZH-EN	32.04	33.80	33.72
Qwen 2.5	DE-EN	32.98	34.18	33.76

LlaMA-2	DE-EN	37.52	34.43	33.37
Mistral	DE-EN	36.61	32.82	33.37

Table 14b. WMT vs Self-Correction Initial vs Final (German-English)

For Qwen 2.5, self-correction training improved the initial translation itself, and the final output remains better than WMT on both language pairs. For LLaMA-2 and Mistral, the pattern is different, their initial translations are worse than their WMT baselines. However, the correction step affected them differently. LLaMA-2's corrections reduced scores further (34.43 - 33.37), while Mistral's corrections improved them (32.82- 33.37).

5.2 Manual Evaluation

To verify whether the metric changes reflect actual improvements, manual evaluation was conducted on 400 samples total. For Qwen 2.5, 100 samples per language pair (200 total) were evaluated. To understand why some models degraded, LLaMA-2 and Mistral were also evaluated on German-English (100 samples each). Each sample was categorized based on whether the initial translation was correct or wrong, whether the model analysed it or left it unchanged, and whether any corrections improved, degraded, or maintained quality.

5.2.1 Results

Tables 15 and 16 show the full breakdown of manual evaluation results for Chinese-English and German-English respectively for Qwen 2.5 model. "Not Analysed" refers to cases where the model's analysis simply states "the translation accurately captures the meaning" without suggesting any changes. "Analysed" refers to cases where the model identified potential issues and attempted correction.

Random sample 100			
Chinese -English			
Total	100		
Base Correct	86	71 Not Analysed	
		15 Analysed	4 Better
			1 Worse
			10 Equal
Base Wrong	14	6 Not Analysed	
		8 Analysed	2 Better
			1 Worse
			5 Equal

Table 15. Manual Evaluation Results (Chinese-English, Qwen 2.5, 100 samples)

Random sample 100			
German -English			
Total	100		
Base Correct	81	20 Not Analysed	
		61 Analysed	7 Better
			10 Worse
			44 Equal
Base Wrong	19	4 Not Analysed	
		15 Analysed	14 Better
			0 Worse
			1 Equal

Table 16. Manual Evaluation Results (German-English, Qwen 2.5, 100 samples)

Most initial translations were already correct: 86% for Chinese-English and 81% for German-English. Since the self-correction model builds on the WMT-fine-tuned model, this reflects the strong translation ability established during the first training stage.

For Chinese-English, Qwen 2.5 responded with “the translation accurately captures the meaning” for 77 of 100 samples, leaving them unchanged. For German-English, it analysed 76 samples. When it identified actual errors, 14 of 15 were fixed successfully (93%) with none made worse. The net result was +4 for Chinese-English (6 better, 2 worse) and +11 for German-English (21 better, 10 worse).

To understand why LLaMA-2 and Mistral showed metric degradation on German-English, manual evaluation was conducted for these models. 100 sentences were randomly sampled from the same 5000-sentence test set used for all evaluations. LLaMA-2 and Mistral were evaluated on identical sentence pairs to enable direct comparison. Tables 17 and 18 show the results.

Random sample 100			
German -English			
Total	100		
Base Correct	77	11 Not Analysed	
		66 Analysed	8 Better
			20 Worse
			38 Equal

Base Wrong	23	0 Not Analysed	
		23 Analysed	7 Better
			4 Worse
			12 Equal

Table 17. Manual Evaluation Results (German-English, LLaMA-2, 100 samples)

Random sample 100			
German -English			
Total	100		
Base Correct	72	8 Not Analysed	
		64 Analysed	12 Better
			20 Worse
			32 Equal
Base Wrong	28	0 Not Analysed	
		28 Analysed	14 Better
			6 Worse
			8 Equal

Table 18. Manual Evaluation Results (German-English, Mistral, 100 samples)

LLaMA-2 showed net degradation of -9 (15 better, 24 worse), while Mistral was neutral at 0 (26 better, 26 worse). Both models analysed more aggressively than Qwen 2.5: LLaMA-2 analysed 89 samples, Mistral analysed 92, compared to Qwen's 76. When fixing actual errors, Mistral succeeded 50% of the time but made 6 worse while LLaMA-2 succeeded only 30% and made 4 worse. Despite using the same training data and training pipeline, the three models showed different outcomes, pointing to model-specific factors.

5.2.2 Examples

This section presents examples from Qwen 2.5 to show how the model behaves in different situations. Using examples from a single model allows for fair comparison across categories. The full breakdown by category with complete sentences is provided in Appendix B. For examples from other models, please refer to Appendix C.

Successful Corrections

Example 1: Terminology Correction (German-English)

(Appendix B: Category 4: Base Wrong + Analysed + Better (Successfully fixed errors))

Source: ...Andilana, die 16 Gaststätten in der Stadt unterhält...

Initial: ...Andilana, which runs 16 hotels in the city...

Corrected: ...Andilana, which runs 16 restaurants in the city...

The German “Gaststätten” means restaurants, not hotels. The model identified and fixed this error.

Example 2: Terminology Correction (Chinese-English)

(Appendix B: Category 4: Base Wrong + Analysed + Better (Successfully fixed errors)

Source: 第二组是所谓”精英控制组...

Initial: The second group is the so-called “elite control group”...

Corrected: The second group, known as “elite controllers”...

“Elite controllers” is the correct medical term for people who naturally suppress HIV. The initial translation made it sound like a research group rather than a type of patient.

Example 3: Omission Correction (Chinese-English)

(Appendix B: Category 2: Base Correct + Analysed + Better (Minor improvements to good translations))

Source: 巴勒斯坦人在以色列工作对双方都有好处...

Initial: Working in Israel is good for both Palestinians and Israelis...

Corrected: Palestinians working in Israel is good for both sides...

The initial translation omitted “Palestinians” as the subject, making it unclear who was working in Israel. The model added this back.

Failed Corrections

Example 4: Number Misinterpretation (Chinese-English)

(Appendix B: Category 3: Base Correct + Analysed + Worse (Overcorrection errors))

Source: ...但这一协议将让纳税人花费700万美元，...

Initial: ...will cost taxpayers $7 million... (correct)

Corrected: ...will cost taxpayers $700,000... (wrong)

The Chinese “700万” means 7 million. The initial translation was correct, but the model wrongly changed it to 700 thousand.

Example 5: Unnecessary Change (German-English)

(Appendix B: Category 3: Base Correct + Analysed + Worse (Overcorrection errors))

Source: ...Goldman Sachs, JP Morgan Chase und Morgan Stanley ihre Hilfen zurückgezahlt.

Initial: ...had already repaid their aid. (correct)

Corrected: ...had already withdrawn their aid. (wrong)

The German “zurückgezahlt” means “repaid”. The initial was correct but the model changed it to “withdrawn”, which has a different meaning.

Example 6: Missed Error (Chinese-English)

(Appendix B: Category 5: Base Wrong + Not Analysed (Missed errors))

Source: ...财阀能获得有利的条件...

Initial: ...zaibatsu receive favorable conditions...

Corrected: ...zaibatsu receive favorable conditions... (unchanged)

The reference uses “chaebol” but the model used “zaibatsu”. The context was about Korean conglomerates, so this was an error. The model did not catch it.

5.2.3 Metrics vs Human Evaluation

One finding from manual evaluation is that automatic metrics do not always reflect genuine translation improvements. In several cases, translations judged as improved during manual evaluation showed decreased BLEU or COMET scores.

To compare automatic metrics with human judgment, sentence-level BLEU and COMET scores were computed for each translation before and after correction. For each sample, the initial translation was scored against the reference, then the corrected translation was scored, and the difference was calculated.

Table 19 shows examples where manual evaluation judged the correction as an improvement, but sentence-level metrics decreased. These examples are from the 100 random samples per language pair, where each sample’s initial and corrected translations were scored against the reference to calculate the change in BLEU and COMET.

Language	**Correction Made**	**BLEU**	**COMET**
ZH-EN	Fixed “following” - “preceded by”	-2.53	-0.0016

ZH-EN	Added “at its core” back	+1.41	-0.0375
DE-EN	Fixed “in” - “at” the plant	-3.83	-0.0236
DE-EN	Fixed “knowledge” - “understanding”	-14.44	-0.0098
DE-EN	Fixed “will be” - “are to be”	-31.56	-0.0211

Table 19. Cases Where Manual Evaluation Disagreed with Automatic Metrics

In these examples, the corrections improved accuracy, fluency, or conveyed the same meaning in different words, but the metrics dropped because the corrected translation differed from the reference. For example, the first row shows a correction from “following the five permanent members” to “preceded by the five permanent members”. Both phrases express the same meaning from different perspectives: the other countries had nuclear weapons before North Korea. The correction is equally valid, but since the reference uses different wording (“joining”), the BLEU score dropped.

This shows a limitation of reference-based metrics. They penalise any difference from the reference, even when the difference is an improvement. Manual evaluation was necessary to confirm that the model was learning meaningful corrections rather than just matching reference phrases.

5.3 Comparison of Dataset Splits

Two dataset configurations were tested. 70-30 (70% error correction examples, 30% clean translations) and 50-50 (equal distribution). Tables 20 and 21 compare performance across these configurations.

Model	**Split**	**BLEU**	**COMET**	**chrF**	**TER**
Qwen 2.5	70-30	**33.72**	0.862	60.43	57.47
	50-50	33.66	**0.864**	**60.75**	**57.36**
LLaMA-3	70-30	33.61	0.861	60.35	57.79
	50-50	**33.67**	**0.862**	**60.39**	**57.63**
LLaMA-2	70-30	30.69	0.855	58.07	60.77
	50-50	**31.27**	**0.857**	**58.38**	**59.49**
Mistral	70-30	30.13	0.852	57.59	61.72
	50-50	**30.80**	**0.854**	**58.05**	**60.72**

Table 20. Comparison of Dataset Splits (Chinese-English)

Model	Split	BLEU	COMET	chrF	TER
Qwen 2.5	70-30	**33.76**	0.841	**58.64**	**55.85**
	50-50	32.06	**0.848**	58.63	61.90
LLaMA-3	70-30	**34.08**	**0.846**	**58.91**	**54.15**
	50-50	30.17	0.819	56.55	65.81
LLaMA-2	70-30	**33.37**	**0.846**	**59.07**	**58.23**
	50-50	32.25	0.843	58.83	61.59
Mistral	70-30	**33.37**	0.843	58.29	**55.48**
	50-50	33.21	**0.844**	**58.33**	55.33

Table 21. Comparison of Dataset Splits (German-English)

For Chinese-English, the 50-50 split performed better or equal for most models. LLaMA-2, LLaMA-3, and Mistral all improved with the balanced split. Qwen showed mixed results, with 70-30 slightly better on BLEU but 50-50 being better on other metrics.

For German-English, the 70-30 split generally worked better. Qwen and LLaMA-3 both showed higher BLEU scores with 70-30, and LLaMA-3 in particular degraded significantly with 50-50. LLaMA-2 and Mistral showed smaller differences between the two configurations.

No single split worked best for all setups.

5.4 Direct Self-Correction Training

An alternative approach where the base models are trained directly on self-correction, skipping the WMT fine-tuning stage. Tables 22 and 23 compare this direct training approach with the two-stage pipeline.

Model	Training Path	BLEU	COMET	chrF	TER
Qwen 2.5	Base - SC Direct	32.82	**0.863**	60.06	58.42
	Base - WMT - SC	**33.72**	0.862	**60.43**	**57.47**
LLaMA-3	Base - SC Direct	31.17	0.860	59.23	62.57
	Base - WMT - SC	**33.67**	**0.862**	**60.39**	**57.63**
LLaMA-2	Base - SC Direct	28.67	0.851	56.45	63.48
	Base - WMT - SC	**31.27**	**0.857**	**58.38**	**59.49**
Mistral	Base - SC Direct	27.50	0.847	55.10	64.23
	Base - WMT - SC	**30.80**	**0.854**	**58.05**	**60.72**

Table 22. Direct versus Two-Stage Training (Chinese-English)

Model	**Training Path**	**BLEU**	**COMET**	**chrF**	**TER**
Qwen 2.5	Base - SC Direct	31.74	0.824	56.27	59.95
	Base - WMT - SC	**33.76**	**0.841**	**58.64**	**55.85**
LLaMA-3	Base - SC Direct	32.72	0.845	58.19	58.84
	Base - WMT - SC	**34.08**	**0.846**	**58.91**	**54.15**
LLaMA-2	Base - SC Direct	26.34	0.741	51.17	76.79
	Base - WMT - SC	**33.37**	**0.846**	**59.07**	**58.23**
Mistral	Base - SC Direct	31.18	0.838	56.84	57.45
	Base - WMT - SC	**33.37**	**0.843**	**58.29**	**55.48**

Table 23. Direct versus Two-Stage Training (German-English)

Both approaches improve from the base model, but the two-stage approach outperformed direct training for all models on both language pairs. The BLEU differences between direct self-correction and two-stage self-correction ranged from 0.90 points (Qwen Chinese-English) to 7.03 points (LLaMA-2 German-English).

LLaMA-2 showed the largest gap on German-English, where direct training resulted in a COMET score of only 0.741 compared to 0.846 with two-stage training. For LLaMA-3 German-English, direct training actually degraded performance below the base model (32.72 vs 34.78 BLEU), while two-stage training maintained it (34.08 BLEU).

These results show that the WMT fine-tuning stage is necessary. It establishes translation ability before the model learns self-correction patterns.

5.5 Summary

Self-correction training improved translation quality for Qwen 2.5 and LLaMA-3 on both language pairs, with smaller or negative effects for Mistral and LLaMA-2. Manual evaluation of 400 samples confirmed that outcomes are model-specific: Qwen 2.5 showed net gains on both language pairs (+4 for Chinese-English, +11 for German-English), while LLaMA-2 showed net degradation (-9) and Mistral was neutral (0) on German-English. The 50-50 dataset split worked better for Chinese-English, while 70-30 was generally better for German-English. The two-stage training approach (Base - WMT - SC) consistently outperformed direct training.

6. DISCUSSION

The results from Chapter 5 showed that the same training approach worked for some models and failed for others. This chapter discusses why that might be the case. It also examines what the manual evaluation revealed about self-correction in practice and compares the findings with existing research.

6.1 Why Some Models Succeeded and Others Failed

The results varied considerably across models despite identical training. On German-English, Qwen 2.5 showed a net improvement of 11 translations, LLaMA-2 showed a net degradation of 9, and Mistral was neutral. On Chinese-English, Qwen 2.5 improved by 4 (net) while other models showed smaller gains. Several factors may explain these differences.

6.1.1 The Role of Pre-training

One likely explanation is what the models were pre-trained on. Qwen 2.5 was built for multilingual use. It supports over 29 languages and was pre-trained on around 18 trillion tokens (Qwen Team, 2024). LLaMA-2 differs significantly. Nearly 90% of its training data is English. German makes up just 0.17% and Chinese only 0.13%. Meta even warned that LLaMA-2 "may not be suitable for use in other languages". (Touvron et al., 2023). Mistral 7B's original publication does not include a detailed breakdown of the languages or sources used in its training data, leaving the exact composition of its corpus unspecified (Jiang et al., 2023).

LLaMA-3 falls somewhere in the middle. Over 5% of its 15 trillion tokens come from more than 30 languages (Meta, 2024). That is better than LLaMA-2, but Meta still says non-English performance may not be as strong.

However, multilingual pre-training does not automatically lead to better base translation scores. On German-English, LLaMA-3 had the highest base BLEU (34.78), followed by LLaMA-2 (32.91) and Mistral (31.28). Qwen 2.5 had the lowest (30.63). Only on Chinese-English did Qwen 2.5 lead, with a base BLEU of 24.89 compared to 23.98 for LLaMA-3.

What multilingual pre-training does seem to help with is the self-correction task itself. Despite having the weakest base score on German-English, Qwen 2.5 showed the strongest self-correction improvements. LLaMA-3, with more multilingual data than LLaMA-2, also performed reasonably well. The models that struggled most with self-correction, LLaMA-2 and Mistral, both have English-heavy pre-training.

Self-correction may require a different capability than basic translation. Understanding how to analyse errors and decide what needs fixing might benefit from exposure to multiple languages during pre-training, even if that exposure does not produce the best raw translation scores.

Another point worth considering is that Qwen 2.5 and Mistral are instruction-tuned, while LLaMA-2 and LLaMA-3 are base models. But this does not seem to be the deciding factor. Qwen 2.5 did best despite being instruction-tuned. Mistral did poorly even though it is also instruction-tuned. Multilingual pre-training appears to matter more than instruction tuning for self-correction.

6.1.2 Baseline Quality and Overcorrection

The models with the best baseline translations were not always the ones that benefited most from self-correction. The models with the highest BLEU scores after WMT fine-tuning were LLaMA-2 (37.52) and Mistral (36.61). Yet these were the same models that performed worse after self-correction training. Qwen 2.5, which had a more modest WMT score of 32.98, showed the largest improvement.

Models that already translate well may have less to gain from self-correction training. If translations are already good, learning to analyse and correct errors might lead the model to make unnecessary changes. The manual evaluation supports this. LLaMA-2 made 20 correct translations worse through overcorrection, compared to 10 for Qwen 2.5.

Examining the numbers more closely shows where the degradation occurs. For LLaMA-2 and Mistral, performance drops before the correction step even begins. Their self-correction models produce initial translations that are worse than their WMT models. LLaMA-2 drops from 37.52 to 34.43 BLEU, and Mistral from 36.61 to 32.82, in the initial translation alone. Qwen 2.5 behaves differently. Its self-correction model produces initial translations that are better than its WMT model (32.98 to 34.18 for German-English).

The correction step also has different effects depending on the model. For Qwen 2.5, corrections slightly reduce scores, but the final output is still above the WMT baseline. For LLaMA-2, corrections reduce scores further, from 34.43 to 33.37. Mistral is an interesting case. Despite producing weaker initial translations, its corrections actually improve the output, going from 32.82 to 33.37 (a gain of 0.55 BLEU). This aligns with the manual evaluation results. LLaMA-2 showed net degradation while Mistral came out neutral overall.

6.1.3 Analysis Format and Language Contamination

There was one key difference between the German-English and Chinese-English training data. The German-English analyses referenced German source text directly, with phrases like "The German 'ausgetauscht' should be 'exchanged' not 'replaced'". The Chinese-English analyses were written entirely in English without any Chinese characters.

This difference may have contributed to the problems in German-English models. During manual evaluation of LLaMA-2, German words appeared in several English output translations. In two cases, analysis text leaked directly into the corrected translation, causing large COMET score drops. These problems did not appear in the Chinese-English evaluations. Having German text in the training analyses may have blurred the boundary for the model between source and target languages.

6.2 The Selective Analysis Behaviour

The self-correction dataset includes examples where the initial translation is already correct, and the analysis simply says "the translation accurately captures the meaning". This teaches the model when to leave translations unchanged. However, models varied considerably in how often they used this conservative response.

For Chinese-English, Qwen 2.5 used "accurately captures" for 77 out of 100 samples and only attempted corrections on 23. It achieved a net improvement of +4. On German-English, Qwen used this response for only 24 out of 100 samples and attempted corrections on 76, yet still achieved +11 net improvement. LLaMA-2 used the conservative response for just 11 out of 100 samples on German-English, attempting corrections on 89, and showed -9 net degradation. Mistral was similar, using it for only 8 out of 100 samples and finishing with 0 net change.

These numbers suggest that analysing too aggressively can lead to worse outcomes. However, Qwen's success on German-English despite analysing 76% of samples shows that the quality of analysis matters more than the frequency. When a model can accurately identify which translations actually contain errors, it can analyse them often and still improve. When it cannot make this distinction reliably, frequent analysis leads to frequent overcorrection.

The difference in analysis rates between language pairs is also notable. Qwen left 77% of Chinese-English translations unchanged but only 24% of German-English translations. A likely explanation relates to the analysis format discussed in Section 6.1.3. The German-English training data included German words in the analysis text, which may have taught models to associate the presence of German with the need to correct something. The Chinese-English analyses, written entirely in English, would not create this association.

6.3 Error Detection Versus Error Correction

Manual evaluation revealed that detecting errors and correcting them are different skills. When Qwen 2.5 correctly identified actual errors in German-English translations, it fixed 14 out of 15 (93%) without introducing new problems. Once it identifies a real error, it usually fixes it successfully.

Mistral and LLaMA-2 were less effective. Mistral fixed 14 of 28 wrong translations (50%) but made 6 worse. LLaMA-2 fixed only 7 of 23 (30%) and made 4 worse. The main problem seems to be detection rather than correction. A model that cannot tell which translations need fixing will overcorrect good translations and miss bad ones.

LLaMA-2 showed some unusual patterns. It produced phrases like "A should be B not B" where both values were the same word. It also hallucinated errors, claiming that sentences were "completely omitted" when they were fully present in the translation. LLaMA-2 seems to have learned the format of error analysis but not when analysis is actually warranted.

Some vocabulary errors appeared in multiple models. Both LLaMA-2 and Mistral mistranslated the German compound "Geländerundgang" (tour of premises). LLaMA-2 translated it as "treasure hunt" while Mistral produced "railway embankment". Mistral also made other vocabulary errors, translating "Gummigeschosse" as "tear gas" instead of "rubber bullets" and "Zähneknirschen" as "toothbrushes" instead of "gnashing of teeth". These examples suggest that certain German vocabulary may be underrepresented in the models' pre-training data.

6.4 Limitations of Automated Metrics

BLEU and COMET scores for Qwen 2.5 dropped slightly between the initial and corrected translations, yet manual evaluation showed net improvements of +4 for Chinese-English and +11 for German-English. Automated metrics suggested the corrections made things worse, while manual evaluation showed the opposite.

Reference-based metrics like BLEU penalise any change that deviates from the reference, even if it improves the translation. A model that paraphrases a sentence or restructures it for clarity will see its scores drop simply because the words no longer match. Table 19 in Chapter 5 includes several examples where corrections were judged as improvements during manual evaluation but received lower BLEU and COMET scores.

For self-correction research, this is an important limitation. If researchers rely only on automated metrics, they may conclude that a self-correction approach does not work when it actually does. Manual evaluation is more time-consuming, but it captures improvements that metrics miss. Even a small sample of 100 translations, as used in this study, can reveal patterns that corpus-level metrics obscure.

6.5 Comparison with Prior Work

Huang et al. (2024) argued that large language models cannot self-correct reasoning without external feedback. When models generate their own feedback, they often fail to identify actual errors or introduce new ones. The results of this thesis partially support that view. LLaMA-2 and Mistral showed the kinds of failures Huang et al. (2024) described, while Qwen 2.5 achieved consistent improvements.

One important difference is the method. Huang et al. (2024) tested prompt-based self-correction at inference time, where models are simply asked to review and improve their outputs. The approach in this thesis embeds the correction capability through fine-tuning on explicit error analyses. The model sees concrete examples of errors and corrections during training, which may help certain architectures learn more reliable patterns.

TEaR needs external evaluation at inference time; this approach handles analysis and correction in a single pass. For Qwen 2.5, this worked well. For LLaMA-2 and Mistral, the lack of external validation meant that errors in the analysis carried through to the corrections.

6.6 Limitations of This Study

This study has several limitations worth noting.

The manual evaluation covered 400 samples in total (100 samples for each of the four evaluations conducted), which is small relative to the 5,000-sentence test sets. Random sampling should provide a representative picture, but patterns in the broader data may have been missed.

The experiments covered only two language pairs, German-English and Chinese-English. Both involve translation into English. It is unclear whether the approach would work for translation out of English or between two non-English languages.

All models were in the 7-8 billion parameter range. Larger models might behave differently and could show better self-correction capabilities due to greater capacity for reasoning.

The self-correction dataset was created using outputs from LLaMA-2. The types of errors are general and occur in all translation models. Future work could test whether using datasets generated from different models leads to different results.

The analysis format also differed between language pairs. Chinese-English analyses were written entirely in English, while German-English analyses included German phrases. All models showed improvement on Chinese-English based on automatic metrics, but results on German-English were mixed. This might point to English-only analyses being more effective. Then again, the base and WMT scores for Chinese-English were lower to begin with, so the two tasks are not directly comparable. A controlled experiment using the same format for both language pairs would help answer this question.

6.7 Summary

This chapter discussed why self-correction training produced different outcomes across models. The next chapter summarises the key findings and outlines directions for future work.

7. CONCLUSION

7.1 Summary of Findings

This thesis explored whether fine-tuning LLMs on explicit error analyses can teach them to self-correct translation errors. The experiments demonstrate that this is achievable for certain models. Qwen 2.5 improved on both language pairs, gaining 1.68 BLEU points on Chinese-English and 0.78 on German-English. Manual evaluation confirmed these gains. Out of 400 samples reviewed, Qwen showed a net improvement of 4 translations on Chinese-English and 11 on German-English. It fixed 93% of the errors it correctly identified. Training in two stages rather than directly on self-correction data also helped, with improvements ranging from 0.9 to 7.0 BLEU points depending on the model and configuration. Still, not every model benefited equally. In German-English, LLaMA-2 actually got worse after self-correction training, and Mistral showed no clear improvement. Pre-training seems to be a key factor in whether self-correction works.

7.2 Contributions

This thesis contributes in several ways. It shows that models can learn to self-correct translations through training, without needing external tools or multiple prompting steps. It also shows that not all models benefit equally from this training. Qwen improved while LLaMA-2 got worse, even though both received the same data. This might explain why other researchers have seen inconsistent results with self-correction. The datasets used in this thesis are available for future work. Each sample includes source text, an initial translation, an error analysis, and a corrected translation. The experiments showed that BLEU and COMET do not always reflect real improvements, so manual evaluation remains necessary.

7.3 Practical Recommendations

Models with multilingual pre-training seem better suited for self-correction. Qwen 2.5, which was trained on many languages, improved consistently. LLaMA-2 and Mistral, which are more English-focused, did not. If the goal is self-correction for non-English translation, starting with a multilingual model is probably a good idea.

Training in two stages also helped. Fine-tuning on translation data first, then on self-correction data, worked better than going straight to self-correction. The first stage builds general translation ability. The second stage refines it.

The ratio of error examples to clean examples in the training data mattered too. Some models did better with a 50-50 split, others with 70-30. Testing both is worth the effort.

Overcorrection was a recurring problem. Some models changed translations that were already fine, which reduced overall quality. Monitoring how often the model changes already correct translations can reveal whether this is happening.

Manual evaluation turned out to be essential. BLEU and COMET scores sometimes went down even when translations actually improved. Reviewing a sample of outputs by hand gives a clearer picture of what is really happening.

Finally, keeping the analysis text in the target language only may help. When German appeared in the training analyses, some models started mixing German into their English outputs. Writing all analyses in English avoided this problem for Chinese-English.

7.4 Future Work

There are several ways future work could build on these findings. One is to test larger models. All models in this thesis were in the 7-8 billion parameter range. Larger models with 70 billion parameters or more might handle self-correction differently, either performing better due to increased capacity or showing the same patterns observed here.

Another direction is to expand to more language pairs. Both pairs in this thesis involved translation into English. It would be useful to know whether the approach works for translation out of English or between two non-English languages.

Accurately detecting errors was the main challenge. Models that could not accurately identify errors either overcorrected or missed problems. Future work could explore ways to improve detection, such as adding confidence scores or external quality checks.

It might also be worth combining training-based and inference-time approaches. A model could be trained on error analyses and prompted to verify its corrections before outputting them. This might reduce overcorrection.

Another possibility is to create model-specific training data. This thesis used a single dataset generated from LLaMA-2 outputs. Training each model on its own errors might produce better results, especially for models that struggled here.

Finally, domain-specific self-correction could be valuable. Training on errors from legal, medical, or technical texts might produce models that are more reliable in those areas.

7.5 Closing Remarks

Self-correction in machine translation remains an open research problem. The reasons why it works for some models but not others are not yet fully understood. However, the results of this thesis show that models can learn to analyse and correct their own translation errors to some extent in a single pass, without relying on external tools or feedback. This indicates that training-based self-correction is a promising direction. Future work can build on these findings to improve the reliability of self-correction across different model architectures and language pairs.

REFERENCES

Alves, D. M., Pombal, J., Guerreiro, N. M., Martins, P. H., Alves, J., Farajian, A., . . . Martins, A. F. (2024). *Tower: An open multilingual large language model for translation-related tasks.* arXiv. https://doi.org/10.48550/arXiv.2402.17733

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Y. Bengio (Ed.), *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015).* Retrieved from https://arxiv.org/abs/1409.0473

Berger, N., Riezler, S., Exel, M., & Huck, M. (2024). Prompting large language models with human error markings for self-correcting machine translation. In C. Scarton (Ed.), *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)* (pp. 636–646). European Association for Machine Translation. Retrieved from https://aclanthology.org/2024.eamt-1.54/

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics, 19*(2), 263-311. Retrieved from https://aclanthology.org/J93-2003/

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., . . . Child, R. (2020). Language models are few-shot learners. In H. Larochelle (Ed.), *Advances in Neural Information Processing Systems. 33*, pp. 1877–1901. Curran Associates. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Chen, P., Guo, Z., Haddow, B., & Heafield, K. (2024). Iterative Translation Refinement with Large Language Models. In C. Scarton (Ed.), *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)* (pp. 181–190). European Association for Machine Translation (EAMT). Retrieved from https://aclanthology.org/2024.eamt-1.17/

Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In D. Wu (Ed.), *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* (pp. 103–111). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-4012

Feng, Z., Chen, R., Zhang, Y., Meng, Z., & Liu, Z. (2024). Ladder: A model-agnostic framework boosting LLM-based machine translation to the next level. In Y. Al-Onaizan (Ed.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 15377–15393). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.860

Feng, Z., Zhang, Y., Li, H., Wu, B., Liao, J., Liu, W., . . . Liu, Z. (2025). TEaR: Improving LLM-based Machine Translation with Systematic Self-Refinement. In L. Chiruzzo (Ed.), *Findings of the Association for Computational Linguistics: NAACL 2025* (pp. 3922–3938). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-naacl.218

Fernandes, P., Deutsch, D., Finkelstein, M., Riley, P., Martins, A., Neubig, G., . . . Firat, O. (2023). The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In P. Koehn (Ed.), *Proceedings of the Eighth Conference on Machine Translation (WMT)* (pp. 1066–1083). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.100

Fomicheva, M., Sun, S., Yankovskaya, L., Blain, F., Guzmán, F., Fishel, M., . . . Specia, L. (2020). Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics, 8*, 539–555. https://doi.org/10.1162/tacl_a_00330

Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Avramidis, E., . . . Martins, A. F. (2022). Results of WMT22 metrics shared task: Stop using BLEU—Neural metrics are better and more robust. In P. Koehn (Ed.), *Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 46–68). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.wmt-1.2

Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Duan, N., & Chen, W. (2024). CRITIC: Large language models can self-correct with tool-interactive critiquing. In B. Kim (Ed.), *Proceedings of the International Conference on Learning Representations*, (pp. 57734–57811). Retrieved from https://proceedings.iclr.cc/paper_files/paper/2024/file/fef126561bbf9d4467dbb8d27334b8fe-Paper-Conference.pdf

Guerreiro, N. M., Alves, D. M., Waldendorf, J., Haddow, B., Birch, A., Colombo, P., & Martins, A. F. (2023). Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics, 11*, 1500–1517.

Guerreiro, N. M., Rei, R., van Stigt, D., Coheur, L., Colombo, P., & Martins, A. F. (2024). xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics, 12*, 979–995. https://doi.org/10.1162/tacl_a_00683

He, Z., Wang, X., Jiao, W., Zhang, Z., Wang, R., Shi, S., & Tu, Z. (2024). Improving machine translation with human feedback: An exploration of quality estimation as a reward model. In K. Duh (Ed.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 8164–8180). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.451

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., . . . Awadalla, H. H. (2023). *How good are GPT models at machine translation? A comprehensive evaluation.* arXiv. https://doi.org/10.48550/arXiv.2302.09210

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., . . . Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In K. Chaudhuri (Ed.), *Proceedings of the 36th International Conference on Machine Learning. 97*, pp. 2790–2799. Proceedings of Machine Learning Research. Retrieved from https://proceedings.mlr.press/v97/houlsby19a.html

Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., . . . Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *Proceedings of the International Conference on Learning Representations.* Retrieved from https://openreview.net/forum?id=nZeVKeeFYf9

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A., Song, X., & Zhou, D. (2024). Large Language Models Cannot Self-Correct Reasoning Yet. In B. Kim (Ed.), *Proceedings of the International Conference on Learning Representations*, (pp. 32808–32824). Retrieved from https://proceedings.iclr.cc/paper_files/paper/2024/file/8b4add8b0aa8749d80a34ca5d941c355-Paper-Conference.pdf

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., . . . Stock, P. (2023). *Mistral 7B.* arXiv. https://doi.org/10.48550/arXiv.2310.06825

Kamoi, R., Zhang, Y., Zhang, N., Han, J., & Zhang, R. (2024). When can LLMs actually correct their own mistakes? A critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics, 12*, 1417–1440. https://doi.org/10.1162/tacl_a_00713

Kocmi, T., & Federmann, C. (2023). Large language models are state-of-the-art evaluators of translation quality. In M. Nurminen (Ed.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation* (pp. 193–203). European Association for Machine Translation. Retrieved from https://aclanthology.org/2023.eamt-1.19/

Koehn, P., & Knowles, R. (2017). Six Challenges for Neural Machine Translation. In T. Luong (Ed.), *Proceedings of the First Workshop on Neural Machine Translation* (pp. 28–39). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3204

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, (pp. 127–133). Retrieved from https://aclanthology.org/N03-1017/

Kumar, A., Zhuang, V., Agarwal, R., Su, Y., Co-Reyes, J., Singh, A., . . . Faust, A. (2025). Training language models to self-correct via reinforcement learning. In Y. Yue (Ed.), *Proceedings of the International Conference on Learning Representations*, (pp. 54523–54549). Retrieved from https://proceedings.iclr.cc/paper_files/paper/2025/file/871ac99fdc5282d0301934d23945ebaa-Paper-Conference.pdf

Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. In C. Zong (Ed.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 4582–4597). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.353

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica, 12*, 455–463. https://doi.org/10.5565/rev/tradumatica.77

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., . . . Clark, P. (2023). Self-refine: Iterative refinement with self-feedback. In A. Oh (Ed.), *Advances in Neural Information*

Processing Systems. 36, pp. 46534–46594. Curran Associates. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf

Meta. (2024, April 18). *Introducing Meta Llama 3: The most capable openly available LLM to date*. Retrieved from Meta AI Blog: https://ai.meta.com/blog/meta-llama-3/

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., . . . Askell, A. (2022). Training language models to follow instructions with human feedback. In S. Koyejo (Ed.), *Advances in Neural Information Processing Systems. 35*, pp. 27730–27744. Curran Associates. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

Pan, L., Saxon, M., Xu, W., Nathani, D., Wang, X., & Wang, W. Y. (2024). Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics, 12*, 484–506. https://doi.org/10.1162/tacl_a_00660

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In P. Isabelle (Ed.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In O. Bojar (Ed.), *Proceedings of the Tenth Workshop on Statistical Machine Translation* (pp. 392–395). Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3049

Post, M. (2018). A call for clarity in reporting BLEU scores. In O. Bojar (Ed.), *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 186–191). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6319

Qwen Team. (2024). *Qwen2.5: A party of foundation models*. Retrieved from Qwen Blog: https://qwenlm.github.io/blog/qwen2.5/

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training.* OpenAI. Retrieved from https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners.* OpenAI. Retrieved from https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Raunak, V., Sharaf, A., Wang, Y., Awadalla, H., & Menezes, A. (2023). Leveraging GPT-4 for automatic translation post-editing. In H. Bouamor (Ed.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 12009–12024). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.804

Rei, R., Stewart, C., Lavie, A., & Farinha, A. (2020). COMET: A neural framework for MT evaluation. In B. Webber (Ed.), *Proceedings of the 2020 Conference on Empirical Methods in Natural*

Language Processing (EMNLP) (pp. 2685–2702). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In K. Erk (Ed.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1715–1725). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1162

Shao, C., Meng, F., Zeng, J., & Zhou, J. (2024). Understanding and addressing the under-translation problem from the perspective of decoding objective. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3800–3814). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.209

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. In A. Oh (Ed.), *Advances in Neural Information Processing Systems. 36*, pp. 8634–8652. Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf

Snover, M., Dorr, B., Makhoul, J., Micciulla, L., & Schwartz, R. (2006). A study of translation edit rate with targeted human annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers* (pp. 223–231). Association for Machine Translation in the Americas. Retrieved from https://aclanthology.org/2006.amta-papers.25/

Stap, D., Hasler, E., Byrne, B., Monz, C., & Tran, K. (2024). The fine-tuning paradox: Boosting translation quality without sacrificing LLM abilities. In L.-W. Ku (Ed.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6189–6206). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.336

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., . . . Cucurull, G. (2023). *LLaMA 2: Open foundation and fine-tuned chat models.* arXiv. https://doi.org/10.48550/arXiv.2307.09288

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. In I. Guyon (Ed.), *Advances in Neural Information Processing Systems. 30*, pp. 5998-6008. Curran Associates. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, Y., Zeng, J., Liu, X., Meng, F., Zhou, J., & Zhang, M. (2024). TasTe: Teaching Large Language Models to Translate through Self-Reflection. In L.-W. Ku, A. Martins, & V. Srikumar (Ed.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6144–6158). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.333

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., . . . Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo (Ed.), *Advances in*

Neural Information Processing Systems. 35, pp. 24824–24837. Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

Welleck, S., Lu, X., West, P., Brahman, F., Shen, T., Khashabi, D., & Choi, Y. (2022). *Generating sequences by learning to self-correct.* arXiv. https://doi.org/10.48550/arXiv.2211.00053

Wu, M., Vu, T.-T., Qu, L., Foster, G., & Haffari, G. (2024). *Adapting large language models for document-level machine translation.* arXiv. https://doi.org/10.48550/arXiv.2401.06468

Xu, H., Kim, Y. J., Sharaf, A., & Awadalla, H. (2024a). A paradigm shift in machine translation: Boosting translation performance of large language models. In B. Kim (Ed.), *Proceedings of the International Conference on Learning Representations*, *2024*, pp. 2747–2767. Retrieved from https://proceedings.iclr.cc/paper_files/paper/2024/file/0b9536e186a77feff516893a5f393f7a-Paper-Conference.pdf

Xu, H., Sharaf, A., Chen, Y., Tan, W., Shen, L., Van Durme, B., . . . Kim, Y. (2024b). Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. *Proceedings of the 41st International Conference on Machine Learning.* Proceedings of Machine Learning Research. Retrieved from https://proceedings.mlr.press/v235/xu24c.html

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., . . . Zhang, H. (2024). *Qwen2.5 technical report.* arXiv. https://doi.org/10.48550/arXiv.2412.15115

Zhu, D., Chen, P., Zhang, M., Haddow, B., Shen, X., & Klakow, D. (2024). Fine-tuning large language models to translate: Will a touch of noisy data in misaligned languages suffice? In Y. Al-Onaizan (Ed.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 388–409). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.24

APPENDIX A: SELF-CORRECTION DATASET SAMPLES

source_de	target_en	predicted_en	error_type	analysis
Diese Plätze können bei Fernverbindungen nicht reserviert werden, sie sind blockiert.	It is not possible to reserve these seats in the coach buses, they are blocked.	These places cannot be reserved for long-distance calls, they are blocked	Semantic Error	Initial translation: These places cannot be reserved for long-distance calls, they are blocked. Analysis: The model completely misunderstood "Fernverbindungen" which means "coach buses" (long-distance transportation) not "long-distance calls" (telecommunications). This is a fundamental domain confusion between transportation and telecommunications. The word "Plätze" means "seats" in this transportation context, not just "places". Corrected translation: "It is not possible to reserve these seats in the coach buses, they are blocked."
Eine Gruppe der Bürgerinitiative "Schweizer Wiese" hat den Bauzaun auf dem Bechtle-Grundstück an der Kurpromenade in Bad Herrenalb verschönert.	A group belonging to the "Schweizer Wiese" citizen's initiative has given the hoarding on the Bechtle plot on the spa promenade in Bad Herrenalb a face-lift.	A group of the Swiss Meadows citizens' initiative has beautified the fence around the Bechtle property on the Kurpromenade in Bad Herrenalb.	Named Entity Error	Initial translation: A group of the Swiss Meadows citizens' initiative has beautified the fence around the Bechtle property on the Kurpromenade in Bad Herrenalb. Analysis: The model translated "Schweizer Wiese" literally as "Swiss Meadows" when it should remain as the proper name "Schweizer Wiese" in quotes. The German "Bauzaun" means "hoarding" (construction barrier) not just "fence". Corrected translation: A group belonging to the "Schweizer Wiese" citizen's initiative has given the hoarding on the Bechtle plot on the spa promenade in Bad Herrenalb a face-lift.

An mehreren Stellen gebe es entsprechende Schmierereien, sagt Jürgen Kohne von der Stadtverwaltung auf Anfrage der WN.	In several places, there are graffiti, says Jürgen Kohne of the city administration on the request of the WN.	There are a number of places where there are such grease points, according to Jürgen Kohne of the city administration, in response to a question from the WN.	Lexical Error	Initial translation: There are a number of places where there are such grease points, according to Jürgen Kohne of the city administration, in response to a question from the WN. Analysis: The German “Schmierereien” means “graffiti” not “grease points” - this is about vandalism, not mechanical issues. The phrase “entsprechende” means “such” or similar graffiti, which was captured. The model also changed the sentence structure unnecessarily and made it more verbose than the original. The context is clearly about urban vandalism, not maintenance issues, but the model misunderstood the key term. Corrected translation: In several places, there are graffiti, says Jürgen Kohne of the city administration on the request of the WN.
Zum Abschluss der Feierlichkeiten aus Anlass seines 160-jährigen Bestehens wird der Liederkranz Dunningen zwei Gottesdienste mit der “Deutschen Messe” von Franz Schubert umrahmen.	To close the celebrations on the occasion of its 160th anniversary, the Liederkranz Dunningen (Dunningen Amateur Choir, or Glee Club, if you like) will perform Franz Schubert’s “Deutsche Messe” (German Mass) to accompany two church services.	To conclude the celebrations for its 160th anniversary, the Liederkranz Dunningen will be holding two services with the ‘Deutsche Messe’ by Franz Schubert.	Omission/ Addition Error	Initial translation: To conclude the celebrations for its 160th anniversary, the Liederkranz Dunningen will be holding two services with the ‘Deutsche Messe’ by Franz Schubert. Analysis: The model omitted the explanatory translation “(Dunningen Amateur Choir, or Glee Club, if you like)” which helps English readers understand what “Liederkranz” means. The phrase “umrahmen” means “to accompany” or “perform to accompany” not just “holding services with”. Corrected translation: To close the celebrations on the occasion of its 160th anniversary, the Liederkranz Dunningen (Dunningen Amateur Choir, or Glee Club, if you like) will perform Franz Schubert’s “Deutsche Messe” (German Mass) to accompany two church services.

Und das merkte man bei ihrer ersten Vorstellung im Badeanzug am Strand des Hilton-Hotels und bei ihren beiden Aufzügen bei der Krönungsfeier.	And that was noted in her first appearance in swimsuit on the beach of Hilton and during her two shows at the evening coronation.	And that was obvious when they first appeared in their bathing suits on the beach at the Hilton Hotel and when they appeared in their two outfits at the coronation.	Grammatical Error	Initial translation: And that was obvious when they first appeared in their bathing suits on the beach at the Hilton Hotel and when they appeared in their two outfits at the coronation. Analysis: The German “ihrer ersten Vorstellung im Badeanzug” should be “her first appearance in swimsuit” not “they first appeared in their bathing suits” - this is about one person (singular feminine), not multiple people. The phrase “bei ihren beiden Aufzügen bei der Krönungsfeier” should be “during her two shows at the evening coronation” not “when they appeared in their two outfits” - “Aufzügen” means performances/shows, not outfits, and “evening coronation” was omitted. The model incorrectly changed from singular to plural throughout. Corrected translation: And that was noted in her first appearance in swimsuit on the beach of Hilton and during her two shows at the evening coronation.
In Petersburg haben wir welche, die Experten sein könnten, aber ihnen wird der Weg versperrt, da es hochrangigerer Experten bedarf, die die Ersteren richtig bewerten und ihnen den Weg freimachen.	Here in St. Petersburg, there are people who could be experts, but they are pushed to the side, because more advanced experts are needed, who will correctly evaluate these experts and give way to them.	In Petersburg, we have experts who could be experts, but they are blocked from doing so because they need to be assessed by higher-ranking experts and given the green light.	Fluency Error	Initial translation: In Petersburg, we have experts who could be experts, but they are blocked from doing so because they need to be assessed by higher-ranking experts and given the green light. Analysis: The model created redundancy with “experts who could be experts” when the German means “people who could be experts”. Corrected translation: Here in St. Petersburg, there are people who could be experts, but they are pushed to the side, because more advanced experts are needed, who will correctly evaluate these experts and give way to them.

Für den Moment können die Behörden Gastarbeitern, die beim ersten Mal durch den Test fallen, nur versprechen, dass sie Zeit für einen Sprachgrundkurs erhalten.	For the moment, the authorities can promise migrant workers who could not pass the test the first time to give time to complete a basic language course.	For the time being, the authorities can only promise that they will receive language courses for the first time.	Incomplete Translation	Initial translation: For the time being, the authorities can only promise that they will receive language courses for the first time. Analysis: The German “Gastarbeitern, die beim ersten Mal durch den Test fallen” (migrant workers who could not pass the test the first time) was completely omitted, losing the key subject and context. The phrase “dass sie Zeit für einen Sprachgrundkurs erhalten” should be “to give time to complete a basic language course” not “that they will receive language courses for the first time.” Corrected translation: For the moment, the authorities can promise migrant workers who could not pass the test the first time to give time to complete a basic language course.

Table A.1 German-English Self-Correction Dataset Sample For Each Error Type

APPENDIX B: MANUAL EVALUATION EXAMPLES BY CATEGORY

Category 1: Base Correct + Not Analysed (Model correctly left good translations as it is)

Lang	**source**	**reference**	**initial_translation**	**analysis**	**corrected_translation**
Chinese-English	**最后**，对伊朗进行军事打击也将标志着一场地区性并且可能是全球性军事和恐怖活动升级与扩大的开始，这对有关各方来说都不啻是一场可怕的梦魇。	Finally, a military attack on Iran would also mark the beginning of a regional, and possibly global, military and terrorist escalation – a nightmare for all concerned.	Finally, a military strike on Iran would mark the beginning of a regional, and possibly global, escalation of military and terrorist activities, which would be a nightmare for all concerned.	the translation accurately captures the meaning	Finally, a military strike on Iran would mark the beginning of a regional, and possibly global, escalation of military and terrorist activities, which would be a nightmare for all concerned.
German - English	Diejenigen, die leistungsstarke Kinder untersuchen, sagen, diese haben oft eine natürliche Affinität für die Schule und einen ihnen innewohnenden Antrieb zum Erfolg.	Those who study high achievers say they often have a natural affinity for school and an innate drive to succeed.	Those who study high-achieving children say that they often have a natural affinity for school and an intrinsic drive to succeed.	the translation accurately captures the meaning	Those who study high-achieving children say that they often have a natural affinity for school and an intrinsic drive to succeed.

Table B.1: Base Correct, Not Analysed

Corrected Translation: Identical to initial - remained unchanged.

Result: correct - Model correctly left acceptable translation as it is.

Category 2: Base Correct + Analysed + Better (Minor improvements to good translations)

Lang	**source**	**reference**	**initial_translation**	**analysis**	**corrected_translation**
Chinese-English	**巴勒斯坦人在以色列工作**对双方都有好处，也比来自遥远国家并在以色列过着孤立的生活的外国	Palestinians working in Israel are good for both sides, and better than the foreigners who come from distant lands and live an isolated life in Israel,	Working in Israel is good for both Palestinians and Israelis, and it is better than having foreigners from distant countries living in isolation in Israel, where they are often	The model adds "and it is better than having foreigners from distant countries living in isolation in Israel, where they are often threatened with expulsion" at the end, which is not in the source text. The model also changes "Palestinians working in Israel" to "Working in Israel is	Palestinians working in Israel is good for both sides, and it is better than having foreigners from distant countries living in isolation in Israel, where they are often threatened with expulsion.

	人好，这些外国人很孤独并且经常处于被开除的威胁之下。	solitary and under constant threat of expulsion.	threatened with expulsion.	good for both Palestinians and Israelis	
German - English	Sondern, damit seine Mitarbeiter die gesamte Fahrzeugtechnik ausprobieren und beherrschen lernen können.	But so that his employees can try out and learn every facit of their vehicle's technology.	Rather, so that his employees can learn to use and master all the vehicle technology.	The German "damit seine Mitarbeiter die gesamte Fahrzeugtechnik ausprobieren und beherrschen lernen können" should be "so that his employees can try out and master all the vehicle technology" not "can learn to use and master all the vehicle technology	Rather, so that his employees can try out and master all the vehicle technology.

Table B.2: Base Correct, Analysed, Better

Chinese-English: Improved - Reference says "Palestinians working in Israel" but initial just says "working in Israel" which could be anybody.

German-English: Improved: Source says "ausprobieren" which means "try out". Initial said "learn to use" but correction fixed it to "try out" which matches the German better.

Category 3: Base Correct + Analysed + Worse (Overcorrection errors)

Lang	**source**	**reference**	**initial_translation**	**analysis**	**corrected_translation**
Chinese-English	但这一协议将让纳税人花费700万美元,并且家庭采暖和空IT奥设备制造商开利公司仍会将1,300个岗位外包给墨西哥。	Yet the deal will cost taxpayers $7 million, and still allow Carrier to outsource 1,300 jobs to Mexico.	But the deal will cost taxpayers $7 million, and Carrier, the home-heating and air-conditioning manufacturer, will still outsource 1,300 jobs to Mexico.	The model changes the amount to $7 million instead of $700,000, which is a 10x exaggeration. The model also adds "home-heating and air-conditioning" to describe Carrier, which is redundant since Carrier is already known to be an HVAC company. The phrase "will still outsource" makes it sound like this is a continuation of past outsourcing, when the original says "will outsource" in the future tense	But the deal will cost taxpayers $700,000, and Carrier, the home-heating and air-conditioning manufacturer, will outsource 1,300 jobs to Mexico.
German - English	Im Juni hatten bereits Goldman Sachs, JP Morgan Chase und Morgan	Goldman Sachs, JP Morgan Chase and Morgan Stanley had already repaid	In June, Goldman Sachs, JP Morgan Chase and Morgan Stanley had	The German "Hilfen zurückgezahlt" should be "aid withdrawn" not "repaid their aid" - this is about withdrawing aid, not repaying it. The phrase "Hilfen zurückgezahlt" literally	In June, Goldman Sachs, JP Morgan Chase and Morgan Stanley had already withdrawn their aid.

	Stanley ihre Hilfen zurückgezahlt.	their support in June.	already repaid their aid.	means “aid returned” or “aid withdrawn” in this context, not “repaid aid	

Table B.3: Base Correct, Analysed, Worse

Chinese-English: Worse - Reference says 7 million dollars and initial had it right. But the analysis wrongly changed it to 700 thousand. Also, the initial matches the source even better then the reference.

German-English: Worse - German “zuruckgezahlt” means “repaid” . The initial was correct but correction wrongly changed to “withdrawn”.

Category 4: Base Wrong + Analysed + Better (Successfully fixed errors)

Lang	source	reference	initial_translation	analysis	corrected_translation
Chinese-English	第二组是所谓”精英控制组“，该组可以在不接受任何治疗的情况下携带艾滋病毒长达10年以上而不发病。	The second group are “elite controllers,” who can carry HIV for ten years or more without treatment and without falling ill.	The second group is the so-called “elite control group,” which can carry HIV for more than 10 years without developing symptoms, even without treatment.	The model changes “elite controllers” to “elite control group,” which are different terms. The model also adds “even without treatment” at the end, which is not in the source and changes the meaning - the source says they can carry the virus for 10+ years without symptoms, not necessarily without treatment	The second group, known as “elite controllers,” can carry the virus for more than 10 years without symptoms, even without treatment.
German - English	Bei einer anderen führenden Gruppe, Andilana, die 16 Gaststätten in der Stadt unterhält (Hostal de la Rita, Market, La Fonda und andere), hat die Eigentümlichkeit des Kalenders zu vorgezogenen Reservierungen, sogar ab Anfang November, geführt.	In another of the most powerful groups, Andilana, with 16 establishments in the city (Hostal de la Rita, Market, La Fonda and others), the peculiarity of the calendar has involved more anticipated reserves, even since the beginning of November.	Another leading group, Andilana, which runs 16 hotels in the city (Hostal de la Rita, Market, La Fonda and others), has the advantage of the calendar, even from the beginning of November.	The German “Andilana, die 16 Gaststätten in der Stadt unterhält” should be “Andilana, which runs 16 restaurants in the city” not “Andilana, which runs 16 hotels in the city	Another leading group, Andilana, which runs 16 restaurants in the city (Hostal de la Rita, Market, La Fonda and others), has the advantage of the calendar, even from the beginning of November.

Table B.4: Base Wrong, Analysed, Better

Chinese-English: Fixed: Elite controllers is the actual medical term used in reference. Initially, it was wrongly translated as an elite control group, which sounds like a research group. Correction fixed it to elite controllers.

German-English: Fixed: German “Gaststatten” means restaurants, not hotels. Initial wrongly said “hotels” but correction fixed it to “restaurants” which is accurate. Reference seems to have missed that part as well.

Category 5: Base Wrong + Not Analysed (Missed errors)

Lang	**source**	**reference**	**initial_translation**	**analysis**	**corrected_translation**
Chinese-English	作为这些资金的交换 —— 这些资金往往被用于耗资巨大的国家工程甚至政治选战 —— 财阀能获得有利的条件，比如廉价银行贷款和有利的监管规则。	In exchange for that money, which is used to finance costly state projects or even political campaigns, the chaebols gain favors, such as cheap bank loans or preferential regulations.	In exchange for these funds, which are often used to finance costly national projects or even political campaigns, zaibatsu receive favorable conditions, such as cheap bank loans and favorable regulatory rules.	the translation accurately captures the meaning	In exchange for these funds, which are often used to finance costly national projects or even political campaigns, zaibatsu receive favorable conditions, such as cheap bank loans and favorable regulatory rules.
German - English	Das mag so manche Handwerker, Ärztinnen, Kassierer und anderweitig erwerbstätige Bürger in diesem Land zu der berechtigten Frage verleitet haben, was eigentlich der Unterschied ist zwischen Arbeit und Sacharbeit, aber sei's drum.	That may have led some craftsmen, doctors, cashiers and other working people in this country to the legitimate question of what the difference really is between work and paperwork, but that's all.	This may have led many craftsmen, doctors, cashiers and other workers in this country to ask themselves what the difference is between work and work, but that is beside the point.	the translation accurately captures the meaning	This may have led many craftsmen, doctors, cashiers and other workers in this country to ask themselves what the difference is between work and work, but that is beside the point.

Table B.5: Base Wrong, Not Analysed

Chinese-English: Missed error - Reference says "chaebol" which is Korean but initial says "zaibatsu" which is Japanese. Wrong terminology for the context. Model did not catch.

German-English: Missed Error - initial said work and work, and model did not catch that.

All examples in this appendix are from Qwen 2.5 manual evaluation (100 samples per language pair). Using examples from a single model allows for fair comparison across categories. For examples from other models, please refer to Appendix C.

Category 6: Base Wrong + Analysed + Equal (Identified but not fixed)

Lang	**source**	**reference**	**initial_translation**	**analysis**	**corrected_translation**
Chinese-English	新的奥巴马政府提出了开支计划，其中赤字超过一万亿美元，这会创下新的纪录。 而这还是在刚离任的布什政府已经创下了有史以来的最高赤字之后。	The new Obama administration is proposing spending plans that would create a record US deficit of more than one trillion dollars – and this coming on top of the outgoing Bush administration’s record deficit.	The new Obama administration’s proposed spending plan, which would exceed $1 trillion in deficits, would set a new record, even though the just-inaugurated Bush administration had already set a record for the highest deficits in history.	The model changes “new Obama administration” to “just-inaugurated Bush administration,” which completely reverses the intended meaning. The model also adds “even though the just-inaugurated Bush administration had already set a record for the highest deficits in history” at the end, which is not in the source text and creates confusion about which administration set the record	The new Obama administration’s proposed spending plan, which would exceed $1 trillion in deficits, would set a new record, even though the just-inaugurated Bush administration had already set a record for the highest deficits in history.
German - English	Beute machen diese Haie trotzdem: Im Magen eines Tieres fanden Forscher die Überreste einer Robbe, ein anderes hatte den Kieferknochen eines jungen Eisbären geschluckt.	However, these sharks still manage to catch prey: In one animal’s stomach, researchers found the remains of a seal; another shark had gobbled down the jawbone of a young polar bear.	Nevertheless, these sharks are making a living: researchers found the remains of a seal in the stomach of one animal, and another had swallowed the jawbone of a young polar bear.	The German “Beute machen diese Haie trotzdem” should be “These sharks are still making a living” not “Nevertheless, these sharks are making a living” - the word “trotzdem” means “despite” not “nevertheless	These sharks are still making a living: researchers found the remains of a seal in the stomach of one animal, and another had swallowed the jawbone of a young polar bear.

Table B.6: Base Wrong, Analysed, Equal

Chinese-English: Not fixed - Reference says “outgoing Bush administration” but initial wrongly says “just-inaugurated Bush” which has the opposite meaning. Analysis caught it but the correction kept the same wrong text.

German-English: Not fixed: According to reference it is supposed to be catch prey but both initial and correction say “making a living” which is wrong. Analysis missed the real error.

APPENDIX C: CODE AND DATA AVAILABILITY

All code, datasets, and evaluation results from this thesis are available at:

GitHub Repository: https://github.com/SandUpt/thesis_project

Training Data:

Dataset	Location
All processed training data	https://github.com/SandUpt/thesis_project/tree/main/data/processed
German-English self-correction dataset	https://github.com/SandUpt/thesis_project/tree/main/data/processed/de_en/self_correction_de_en
Chinese-English self-correction dataset	https://github.com/SandUpt/thesis_project/tree/main/data/processed/zh_en/self_correction_zh_en

Evaluation / Test Data

Dataset	Location
Test sets (both language pairs)	https://github.com/SandUpt/thesis_project/tree/main/data/evaluation_sets

Manual Evaluation Results

File	Description	Location
File: Multiple models Manual Evaluation Category Examples.xlsx	Compiled Chinese-English examples from all models (LLaMA-2, LLaMA-3, Mistral, Qwen 2.5) across all evaluation categories.	https://github.com/SandUpt/thesis_project/tree/main/evaluations/sampled_data_100_manual_evaluations
Randomly sampled evaluation files	100 randomly sampled outputs from Qwen (ZH-EN, DE-EN), LLaMA-2 (DE-EN), and Mistral (DE-EN)	https://github.com/SandUpt/thesis_project/tree/main/evaluations/sampled_data_100_manual_evaluations

SELBSTSTÄNDIGKEITSERKLÄRUNG

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne Hilfe Dritter und ohne Zuhilfenahme anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

14.12.2025 Sandeep

Ort, Datum Unterschrift