Investigating Attribute-Controlled Translation with Large Language Models

Bachelor's Thesis of

Sijia Huang

Artificial Intelligence for Language Technologies (AI4LT) Lab
Institute for Anthropomatics and Robotics (IAR)
KIT Department of Informatics

Reviewer:	Prof. Dr. Jan Niehues
Second reviewer:	Prof. Dr.-Ing. Rainer Stiefelhagen
Advisor:	M.Sc. Danni Liu

30. November 2024 – 31. March 2025

Karlsruher Institut für Technologie
Fakultät für Informatik
Postfach 6980
76128 Karlsruhe

Karlsruhe, 31.3.2025

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

PLACE, DATE

Sijia Huang

(Sijia Huang)

Abstract

Translation procedures involving attribute control face significant challenges, particularly regarding ambiguity and multiple interpretations. While attributes such as tone and sentiment are inherently subjective, others like gender and formality vary across languages. Additionally, fine-grained attribute data remains scarce.

This study examines Large Language Models' (LLMs) potential for context-aware translation. We conducted experiments comparing standard translation, LLM-based translation, and post-editing approaches to assess LLMs' impact on translation quality. Additionally, we evaluate LLMs' ability to detect missing attributes, particularly in cases of ambiguous or unambiguous gender references.

LLM-based translation includes zero-shot and few-shot setups, incorporating clear instructions and desired attributes in the prompt. The few-shot setup provides additional translation pairs as examples, allowing LLMs to learn patterns and improve performance. In the multilingual translation module, we follow the few-shot approach but use third-language examples for prompting. Post-editing involves refining candidate translations.

Our findings show that LLMs enhance both translation quality and attribute control, except in the counterfactual gender dataset, where standard translation achieves better quality control. However, LLMs consistently outperform standard translation in attribute control. Few-shot setups surpass zero-shot in both quality and attribute control, except for gender accuracy in the counterfactual dataset, possibly due to synonym mismatches. In post-editing, LLMs significantly improve attribute control while maintaining original translation quality. In multilingual translation, we observe a trade-off between quality and attribute control. Lastly, LLMs demonstrate limited accuracy in detecting missing attributes in the source text.

Kurzfassung

Übersetzungsverfahren mit Attributkontrolle stehen vor erheblichen Herausforderungen, insbesondere im Hinblick auf Mehrdeutigkeiten und Mehrfachinterpretationen. Während Attribute wie Ton und Stimmung grundsätzlich subjektiv sind, variieren andere wie Geschlecht und Formalität zwischen Sprachen. Zudem sind detaillierte Attributdaten nach wie vor rar.

Diese Studie untersucht das Potenzial von Large Language Models (LLMs) für kontextsensitive Übersetzung. Wir haben Experimente durchgeführt, in denen wir Standardübersetzung, LLM-basierte Übersetzung und Post-Editing-Ansätze verglichen haben, um den Einfluss von LLMs auf die Übersetzungsqualität zu bewerten. Darüber hinaus bewerten wir die Fähigkeit von LLMs, fehlende Attribute zu erkennen, insbesondere bei mehrdeutigen oder eindeutigen Geschlechtsreferenzen.

LLM-basierte Übersetzung umfasst Zero-Shot- und Few-Shot-Setups, die klare Anweisungen und gewünschte Attribute in die Eingabeaufforderung integrieren. Das Few-Shot-Setup bietet zusätzliche Übersetzungspaare als Beispiele, sodass LLMs Muster lernen und ihre Leistung verbessern können. Im mehrsprachigen Übersetzungsmodul verfolgen wir den Few-Shot-Ansatz, verwenden jedoch Beispiele aus Drittsprachen für die Eingabeaufforderung. Post-Editing umfasst die Verfeinerung von Übersetzungskandidaten.

Unsere Ergebnisse zeigen, dass LLMs sowohl die Übersetzungsqualität als auch die Attributkontrolle verbessern, mit Ausnahme des kontrafaktischen Gender-Datensatzes, wo die Standardübersetzung eine bessere Qualitätskontrolle erreicht. LLMs übertreffen die Standardübersetzung jedoch durchweg bei der Attributkontrolle. Few-Shot-Setups übertreffen Zero-Shots sowohl in der Qualität als auch in der Attributkontrolle, mit Ausnahme der Geschlechtsgenauigkeit im kontrafaktischen Datensatz, möglicherweise aufgrund von Synonym-Fehlpaarungen. Im Post-Editing verbessern LLMs die Attributkontrolle deutlich, während die ursprüngliche Übersetzungsqualität erhalten bleibt. Bei mehrsprachigen Übersetzungen beobachten wir einen Kompromiss zwischen Qualität und Attributkontrolle. Schließlich zeigen LLMs eine eingeschränkte Genauigkeit bei der Erkennung fehlender Attribute im Quelltext.

Contents

1. Introduction

1.1. Motivation

Attribute-controlled translation involves generating translations that respect specific characteristics or constraints, such as formality, gender, tone, sentiment, or domain-specific vocabulary. While this approach offers a more customized translation experience, several challenges emerge. The most fundamental challenge is ambiguity and multiple interpretations involved in translation procedure. While attributes like tone or sentiment can be subjective, attributes like gender and formality are language dependent. Different languages express these attributes in varied ways. Managing this ambiguity while preserving the original intent can lead to inconsistent or inaccurate translations. The other noticeable challenge we encounter is data scarcity for fine-grained attributes. The lack of high-quality, attribute-annotated data poses a challenge for training models to handle nuanced translation requests.

One of the solutions to the above challenges is the integration of large language model (LLMs) into translation workflows. We will mainly focus on attribute-control translation tasks in gender and formality. We want to investigate whether LLM can offer the potential for more precise and context-aware translations and perform well in terms of attribute-control. We will run experiments on standard translation, LLM-based translation and post-editing tasks to observe the impact of LLM in improvement of translation.

1.2. Research Questions

Our research question focuses on conducting an analysis of quality and attribute control of translation from large language models (LLMs).

RQ 1: How well can current state-of-the-art LLMs (e.g., Llama 3.1) achieve attribute-controlled translation into diverse target languages? How much does the performance differ under zero-shot and few-shot setups?
By providing additional examples in prompts, LLMs may identify grammatical patterns and adjust translations based on learned knowledge. Comparing zero-shot and few-shot setups allows us to assess the model's ability to handle different scenarios.

RQ 2: To what extent can LLM-based post-editing improve the output of standard translation models (in both quality and attribute control)?
With specific instructions in the prompt, LLMs can focus on attribute control, potentially improving translation quality. The study aims to evaluate performance in both quality and attribute control under this setup.

RQ 3: To what extent can attribute-controlled translation examples from different languages help?

This research evaluates how the inclusion of few-shot translation examples from different languages, especially those from the same language root, can help LLMs detect attribute control changes and apply them to new translation tasks.

RQ 4: How can we detect from the input sentence alone whether the model needs additional attribute information?

The study explores whether LLMs can identify the necessity for additional attribute information before translation. The model may be asked to rate the need for additional attribute information on a 0-10 scale, providing insights into its awareness and confidence in its translation task.

2. Background and Related Works

This chapter introduces the concepts presented in this thesis. The first part will focus on NLP models. We start with transformer model and transformer-based architectures. Then we move to state of the art models for translation such as NLLB and LLM-based MT models. Next we will explain the attributed controlled translation and challenges involved. The second part is related work in attributed controlled translation with LLMs. We highlight the main differences between this thesis and prior works.

2.1. Language Modeling

2.1.1. Formal Definition

A formal definition of language modeling task is "to learn the joint probability function of sequences of words in a language" (Bengio et al., 2003).

$$P(x) = \prod_{i=1}^{n} P(x_i|x_1, ..., x_{i-1}) \tag{2.1}$$

x is a sequence which is composed of words $\{x_1, x_2, ...x_n\}$.
$P(x)$ is the probability of generating this sequence.
$P(x) = P(x_1, x_2, \ldots, x_n)$ is factorized as $P(x_1) \times P(x_2 \mid x_1) \times \cdots \times P(x_n \mid x_1, \ldots, x_{n-1})$.

This allows models to predict the probability distribution of the next token given previous tokens, which is the fundamental operation underlying text generation, completion, and other NLP tasks.

2.1.2. Transformer Model

The Transformer model (Vaswani et al., 2017) is built based on the sophisticated architecture where multi-head attention plays a significant role. Other models like RNNs (Recurrent neural networks) (Werbos, 1990) process data sequentially, one token at a time, carrying a hidden state forward, which causes a bottleneck for efficiency, especially for long sequences. The mutli-attention mechanism enables Transformer to perform parallelization by processing the entire sequence at once.

To start with, a text input is tokenized, mapped into a sequence of embedding tokens, and then encoded using positional encoding. Positional encoding injects the information of relative or absolute positions of the tokens, so Transformer can be aware of the order of the sequence. Then the tokens are passed into encoder.

The **encoder** is composed of stacks of layers. Each layer includes two sub-layers: a multi-head attention mechanism, followed by a fully connected feed-forward network. After each sub-layer a normalization is applied. After multiple encoder layers, the final representation of each token is passed to the decoder.

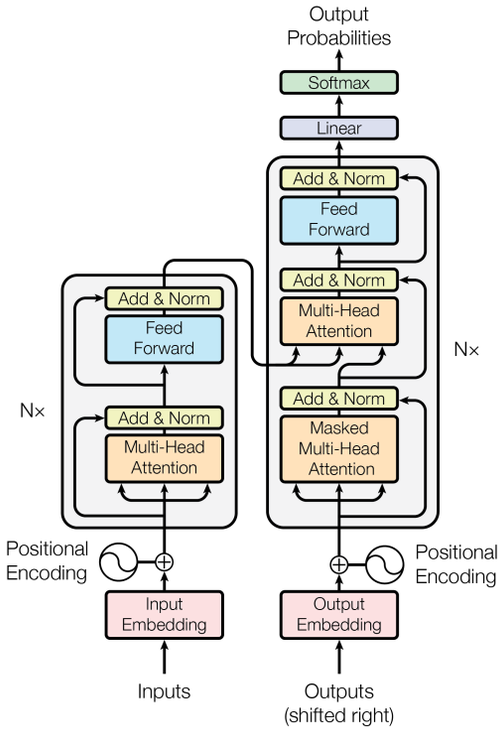


Figure 2.1.: Multi-head attention

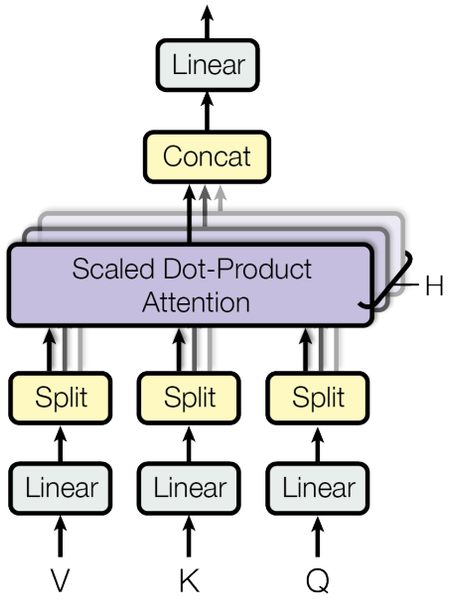
Like the encoder, **decoder** has multiple layers, but with an additional attention mechanism. Decoder started with positional encoding and then enter into a layer composed with three sub-layers: Masked multi-attention, multi-head attention and feed-forward network. After each sub-layer normalization is applied. Masked self-attention is autoregressive, which means that the decoder can only attend to past tokens in the sequence to prevent "seeing the future." The multi-head attention mechanism in decoder performs over the encoder's output, which helps the decoder attend to relevant encoder states. The output of the final layer will be converted into predicted next-token probabilities through learned linear transformation and softmax function.

Multi-head attention mechanism divides its processing into multiple parallel heads. Each head independently performs its own attention computation on different parts of the input representation. These separate attention outputs are then concatenated together and linearly transformed before moving to the next layer. This multi-head approach allows the model to simultaneously attend to information from different representation spaces at different positions, capturing various aspects of the input data in parallel.

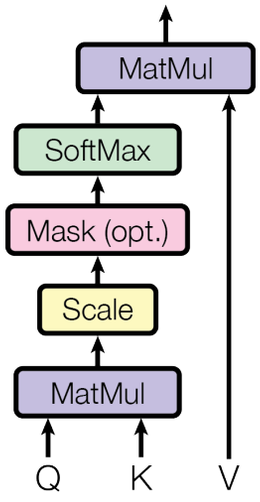
For each layer in multi-head attention we map a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key.

In each layer, we have a Scaled Dot-Product Attention. The input consists of queries and keys of dimension d_k, and values of dimension d_v. We compute the dot products of the query with all keys, divide each by $\sqrt{d_k}$, and apply a softmax function to obtain the weights on the values.

In practice, we compute the attention function on a set of queries simultaneously, packed together into a matrix Q. The keys and values are also packed together into matrices K and V. We compute the matrix of outputs as:



(a) Multi-head attention



(b) Scaled dot-product attention

Figure 2.2.: Comparison of Multi-head and Scaled Dot-Product Attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.2}$$

Feed-forward networks (FFN) consists of two linear transformations with a ReLU (Rectified Linear Unit) activation in between. The input vector is denoted as x, while W_1 and W_2 represent the weight matrices for the hidden and output layers, respectively. Similarly, b_1 and b_2 are the corresponding bias terms for these layers. The function $\max(0, \cdot)$ represents the ReLU activation, which is applied to the hidden layer to introduce non-linearity into the network.

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{2.3}$$

2.1.3. Architectures

2.1.3.1. Encoder-only

An encoder-only model focuses on processing the input data to create a contextual representation of it. It encodes the entire input before producing any output. Due to the lack of decoding mechanism, it cannot convert its representations back into text in a target language. Thus it is not suitable for translation tasks.

An example of Encoder-only model is BERT (Devlin, 2018)(Bidirectional encoder representations from transformers). BERT learns to represent text as a sequence of vectors using self-supervised learning. BERT is trained by masked token prediction and next sentence prediction. As a result of this training process, BERT learns contextual, latent representations of tokens in their context.

2.1.3.2. Encoder-Decoder

Encoder-Decoder model is particularly effective for tasks like machine translation, where understanding the source language is crucial for producing an accurate and fluent translation in the target language.

T5 (Text-to-Text Transfer Transformer) (Raffel et al., 2020) is an example of Encoder-Decoder model. It use a standard encoder-decoder Transformer (Vaswani et al., 2017). The underlying idea is to formulate every text processing problem into a "text-to-text" problem, i.e. taking text as input and producing new text as output. Every task, including translation, question answering, and classification, will be cast as text and fed into the model. Then the model will generate the target text as output.

2.1.3.3. Decoder-only

Decoder-only models are designed to generate output sequences based solely on the preceding context and are trained using supervised learning techniques on large corpora of text data. The training objective often involves maximizing the likelihood of the correct token given the previous tokens, commonly using loss functions such as cross-entropy loss. The examples of Decoder-only model include LLaMA.

LLaMA 3.2 introduces a family of language models with varying capabilities: text-only models at 1B and 3B parameters, and multimodal models handling both text and images at 11B and 90B parameters. These models were derived from LLaMA 3.1(Dubey et al., 2024), which was originally trained with 405B parameters and 126 layers. The development of LLaMA 3.2 included a multi-stage training process starting with LLaMA 3.1 as the base. Image adapters were added to the larger models for multimodal tasks, and the smaller models were optimized through structured pruning and knowledge distillation to retain performance while reducing model size. In post-training, techniques such as supervised fine-tuning (SFT), rejection sampling (RS), and direct preference optimization (DPO) were applied to further align these models for robust, real-world applications.

2.2. Machine Translation

2.2.1. Formal Definition

Machine Translation (MT) is defined as "The automatic translation of text from one human language into another" (Kenny, 2018) or "Conceived as computational systems that translate texts from one language to another" by NLLB team (Costa-jussà, Cross, et al., 2022). It has evolved from Statistical Machine Translation(SMT) (Lopez, 2008) towards encoder-decoder Neural Machine Translation (NMT) (Sutskever, Vinyals, and Le, 2014; Bahdanau, Cho, and Bengio, 2014). In recent years there exists more development of LLM-based MT models.

NMT is defined as "an end-to-end learning approach for automated translation, with the potential to overcome many of the weaknesses of conventional phrase-based translation systems." (Wu et al., 2016). Traditional machine translations such as Deep Neural Networks (DNNs) work well with large labeled training sets, but they cannot be used to map sequences to sequences. NMT typically consists of two recurrent neural networks (RNNs), one to consume the input text sequence and one to generate translated output text.

In NMT we use this method in 2.4 is for tasks like language modeling and text generation, where the likelihood of the next word is determined by the previous words and relevant context

c. The probability of a word sequence x given a condition c, denoted as $P(x, c)$, is calculated by multiplying the probability of each word appearing in order, conditioned on the preceding words and c.

$$P(x, c) = \prod_{i=1}^{n} P(x_i \mid x_1, ..., x_{i-1}, c) \tag{2.4}$$

x is a sequence which is composed of words $\{x_1, x_2, ...x_n\}$.
c is a condition.
$P(x, c)$ is the probability of generating this sequence with c as condition.
$P(x, c) = P(x_1, x_2, \ldots, x_n, c)$ is factorized as $P(x_1, c) \times P(x_2 \mid x_1, c) \times \cdots \times P(x_n \mid x_1, \ldots, x_{n-1}, c)$.

2.2.2. State-of-the-Art Models

2.2.2.1. Dedicated Models

NLLB-200 (Costa-jussà, Cross, et al., 2022) (No Language Left Behind) currently has 1.3B, 3.3B, distilled 600M and distilled 1.3B models. NLLB-200 uses an encoder-decoder model for translation over 200 languages. Since some low-resource languages (Joshi et al., 2020) lack data availability, a many-to-many multilingual human-translated dataset FLORES-200 (Costa-jussà, Cross, et al., 2022) was built based on FLORES-101 (Goyal et al., 2022), whose coverage capped at 100 languages. Flores-200 covers 204 languages and consists of translations from 842 distinct web articles, totaling 3001 sentences. These sentences are divided into three splits: dev, devtest, and test.

To collect highly accurate parallel texts in more languages, sentence encoder LASER3 (Costa-jussà, Cross, et al., 2022) (Language-agnostic sentence representations) was developed from the earlier version LASER (Heffernan, Çelebi, and Schwenk, 2022). LASER3 used Transformer model and teacher-student training and language-group-specific encoders to boost performance. This enables LASER3 to scale language coverage identifying aligned bitext for 148 languages.

2.2.2.2. LLM-based MT models

The examples of LLM-based MT model include Tower (D. M. Alves et al., 2024), ALMA (Xu et al., 2023) and Aya-101 (Üstün et al., 2024).

TOWER was extended from Llama 2 (Touvron et al., 2023) and through continued pretraining on a multilingual mixture of monolingual and parallel data, creating TOWERBASE model. The training dataset comprises 20 billion tokens for 10 languages : English (en), German (de), French (fr), Dutch (nl), Italian (it), Spanish (es), Portuguese (pt), Korean (ko), Russian (ru), and Chinese (zh). Then the dataset to specialize LLMs for translation-related tasks called TOWERBLOCKS is created. TOWERBLOCKS focuses on both data diversity and quality. It collects records from various existing datasets, reformulating them into question-answer pairs, emphasizing zero-shot and few-shot instructions to enhance multilingual understanding. For quality, the dataset uses human-annotated records, avoids data from 2023 onwards, and filters out low-quality translations and translationese. TOWERBLOCKS is used for finetuning TOWERBASE model, creating TOWERINSTRUCT model.

ALMA (Advanced Language Model-based trAnslator) is based on LLaMA-2 7B model. The development of ALMA model comprises two stages: continuous monolingual data fine-tuning and high-quality parallel data fine-tuning. To address conventional LLMs' bias towards English-dominated corpora, ALMA incorporated monolingual data from non-English languages.

Aya-101 covers 101 languages, 53% of which are lower-resourced. It is built on mT5 model (Xue et al., 2021). The mT5 model consists of 13 billion parameters, with 1 billion parameters allocated to token embeddings. Aya-101 is fine-tuned using the Adafactor optimizer (Shazeer and Stern, 2018), with a learning rate of 3×10^{-4} and a batch size of 256.

2.2.3. Attribute-controlled translation

Attribute-controlled translation (ACT) (Sarti et al., 2023) is a subtask of machine translation that involves controlling stylistic or linguistic attributes (such as formality and gender) in translation outputs. ACT takes three inputs: a sentence x , a condition c and a desired target attribute a. The goal is to produce a translation y that is chosen from the highest probability that aligns with the specified attribute. This can be formulated as:

$$P(x, c, a) = \prod_{i=1}^{n} P(x_i | x_1, ..., x_{i-1}, c, a) \tag{2.5}$$

x is a sequence which is composed of words $\{x_1, x_2, ...x_n\}$.
$P(x, c, a)$ is the probability of generating output y with condition c and attribute a.
$P(x, c, a) = P(x_1, x_2, \ldots, x_n, c, a)$ is factorized as $P(x_1, c, a) \times P(x_2 \mid x_1, c, a) \times \cdots \times P(x_n \mid x_1, \ldots, x_{n-1}, c, a)$.

2.3. Related Work

	Related work	Focus	Method	Model	Relevance to this thesis
1	Fine-grained Gender Control with LLMs (Section 2.3.1)	Gender-of-Entity (GoE) prompting for LLMs	Checks LLM's ability to derive correct gender of ambiguous entities	Llama 2 70B, ChatGPT 3.5	RQ4: Ambiguity detection; LLM as Gender Evaluator metric.
2	Generating Gender Alternatives in Machine Translation (Section 2.3.2)	Generating gender alternatives for ambiguous entities	Creates a mapping from gender-ambiguous entities to gender structures	M2M 1.2B, GPT-3.5-turbo	RQ1 and RQ2: Translation with ambiguous entities
3	Gender-specific Machine Translation with Large Language Models (Section 2.3.3)	Using in-context examples to translate gender-neutral source to gender-specific targets	Runs LLM in few-shot to output two sentences with different genders	NLLB 3B, Llama-2 70B	RQ1: Comparison of standard translation and LLM performance
4	Leveraging GPT-4 for Automatic Translation Post-Editing (Section 2.3.4)	Using LLM for translation post-editing to improve quality and remove errors	Post-editing with or without Chain of Thought (CoT)	GPT-4, GPT-3.5-turbo	RQ2: Comparison of standard translation and Post-editing models
5	Transferability of Attribute Controllers on Pretrained Multilingual Translation Models (Section 2.3.5)	Assessing the transferability of attribute controllers across languages	Trains classifiers on decoder activations to adjust model activations	NLLB-200, Transformer trained from scratch (on OPUS-100)	RQ3: Use examples from a third language to enhance translation quality
6	Enhanced Prompting for Attribute-Controlled Translation (Section 2.3.6)	Enhances Attribute-Controlled Translation using LLMs in few-shot and zero-shot	Semantic Similarity Retrieval and Attribute Marking	XGLM, BLOOM, GPT-NEOX	RQ1: Comparison of standard translation and LLM performance

Table 2.1.: Summary of Related Work

2.3.1. Fine-grained Gender Control with LLMs

Lee(Lee et al., 2024) investigated the Gender-of-Entity(GoE) (Lee et al., 2024) prompting method for LLMs where LLMs are explicitly instructed to translate the source text with additional entity-level gender information. More specifically, the prompt includes "Gender Annotation: for `[ENT_1]`, use `[GENDER_1]`; ...; for `[ENT_n]`, use `[GENDER_n]`" where `ENT_i` refers to the *i*-th entity and `GENDER_i` refers to the *i*-th entity's correct gender.

The input sentences are divided into four various scenarios: sentences with single ambiguously gendered entity, multiple ambiguously gendered entities, mixed gendered entities, and complex unambiguous entities. For assessment, single ambiguous entity is selected from MuST-SHE (Bentivogli et al., 2020), multiple ambiguous entities are from GATE (Rarrick et al., 2023), mixed entities are from WinoMT (Stanovsky, Smith, and Zettlemoyer, 2019) and dataset from Saunder (Saunders, Sallis, and Byrne, 2020), and complex unambiguous entities are from MT-GenEval (Currey et al., 2022). The author uses NLLB as comparison to investigate the performance of LLMs Llama 2 70B and ChatGPT 3.5.

The author further creates three baseline methods based on NLLB-200 600M: gender prefixing, gender-specific fine-tuning (FT), and inference-time classifier guidance (CG). For LLM models, author executes experiment with baseline and GoE prompting setup. For scenario with mixed entities, it is discovered that the specified gender of ambiguous gender entity in prompt will interfere the gender of the other entity that are gender unambiguous. The author thus created two GoE prompts: `GoE_amb` and `GoE_full`, where the former specifies only the gender of the ambiguous gender entity, while the latter gives out all entity genders. For scenario with complex entities, author applies Gender-Aware Contrastive Learning(GACL) method on NLLB to further improve gender-debiasing. For scenario with complex entities which runs on MT-GenEval's contextual dataset, to solve the lack of entity annotation, author uses the Spacy 3 dependency parser to extract the noun phrase of the second sentence while using the gendered word list (Zhao et al., 2017) to extract the gender of the entity in the first sentence(the first sentence is context that should be used to infer the gender of the entity).

In this work, because the automated gender accuracy metric is dependent on the annotated gender terms, which poses problems with synonyms or grammatical structures, the author also evaluated the gender-controlled performance by using LLM as Gender Evaluators (LGE).

The key difference between this study and my work lies in the Gender-of-Entity (GoE) prompting method. Unlike GoE, my approach doesn't extract nouns or map entities to specific genders. Instead, my prompts use general wording to indicate the presence of a human and the desired feminine or masculine gender form. This aims for a broader gender understanding by the LLM rather than an entity-specific one.

Specifically, when dealing with complex entities in MT-GenEval's contextual dataset (where gender isn't explicitly provided), my prompts instruct the LLM to infer the correct gender form logically from the surrounding context. This tests the LLM's ability to deduce gender from context.

Furthermore, my experiments use the standard NLLB baseline without any fine-tuning, allowing for a direct performance comparison between the LLM and a traditional NMT model.

2.3.2. Generating Gender Alternatives in Machine Translation

When the MT system is not able to disambiguate gender through context, this study (Garg et al., 2024) suggests providing multiple translation alternatives that cover all valid gender choices. The entity-level alternatives are grouped into a single structured translation with embedded gender structures.

The study uses test sets from GATE's gender-ambiguous dataset and MT-GenEval's contextual dataset where the gender can be inferred from the sentence context. Theses datasets are later post-edited to include marked entities and gender-marked head words. Head word is representative of words that are referring to the same entity. For training data, the study uses train sets from Europarl (Koehn, 2005), WikiTitles (Tiedemann, 2012), and WikiMatrix (Schwenk et al., 2021) corpora. The train data is split into G-Tag and G-Trans, the former contains gender-marked headwords and the latter contains gender-ambiguous entities in the source sentences, gender structures in the translations and gender alignments.

It also developed a semi-supervised approach that leverages pre-trained MT models (fine-tuned M2M 1.2B (Fan et al., 2020) model using fairseq (Ott et al., 2019)) or LLMs (gpt-3.5-turbo model) for data augmentation.

This study addresses attribute ambiguity by generating all possible translations, whereas my thesis aims to produce a single, definitive translation. This single-output approach is more practical for users unfamiliar with the target language, eliminating the need to choose among alternatives. Furthermore, while this study focuses on data augmentation and model training, my thesis centers on analyzing the inherent behavior of existing LLM models in response to varied prompt designs.

2.3.3. Gender-specific Machine Translation with Large Language Models

This study (Sánchez et al., 2023) focuses on using in-context examples (ICEs) to translate from a gender-neutral source sentence to two gender-specific target sentences. It evaluates translation in gender control and quality control. For gender control, it uses coreference resolution accuracy and for quality control it uses BLEU. It uses FLoRes to prove the reliance on coreference resolution of the gender-specific translation method.

This study uses MULTILINGUAL HOLISTIC BIAS (MHB) (Costa-jussà, Andrews, et al., 2023) dataset as gender-focused dataset, BUG's (Levy, Lazar, and Stanovsky, 2021) gold set for gender bias analysis, and FLoRes devtest set as general translation dataset. It uses NLLB 3B model and Llama-2 70B model with both ICEs and standard MT template to execute translation tasks.

A key difference from this study is that my thesis focuses on single-output generation per input, rather than producing both masculine and feminine translations. The study's dual-gender outputs allow for an investigation into coreference resolution using gender-specific (MHB) and general (FLoRes) datasets by analyzing BLEU score differences, which is not a focus of my thesis.

2.3.4. Leveraging GPT-4 for Automatic Translation Post-Editing

Raunak(Raunak et al., 2023) demonstrated that GPT-4 (Achiam et al., 2023) is adept at translation post-editing, producing meaningful edits to translations that help improve its general quality and remove major errors in the text. This study focuses on four research questions: Nature of the Post-Edited Translation, General Quality Improvements, Edits On Human Annotated Error Spans and Trustworthiness of the Proposed Edits.

The experiment used WMT-22 General MT translation task datasets (Kocmi et al., 2022) as well as WMT-20 and WMT-21 News translation task submissions annotated with MQM (Multidimensional Quality Metrics Framework) errors (Freitag et al., 2021). For LLMs, author uses GPT-4 and gpt-3.5-turbo in the experiments. For initial translation, the author uses Microsoft-Translator and other NMT systems. The post-edit prompts are set under three settings: (i) post-editing with a Chain of Thought (CoT), (ii) post-editing without CoT and

(iii) post-editing with Structured-CoT (SCoT). Apart from instructions for translation task, CoT here specifically means that LLM is asked to give proposed improvement steps and then provide the improved translations. Structured-CoT (SCoT) is defined as CoT in the form of an MQM annotation.

For RQ1, the author investigates whether LLM is translating directly from source sentence even though it produces steps of proposed improvement. For RQ2 the author investigates the quality of improved translation from post-editing tasks. For RQ3 the author investigates whether LLM is capable of discovering the errors in text and modifying them. For RQ4 the author investigates whether proposed improvement exists in the improved translation.

In contrast to this study, my thesis does not employ Chain-of-Thought (CoT) prompting. As this study acknowledges the potential for LLMs to disregard provided CoT and the risk of CoT diverting focus from translation quality, my work prioritizes a post-editing task with minimal distractions, excluding CoT. Furthermore, my thesis does not evaluate the LLM's ability to identify and correct specific errors, but rather focuses on improvements in overall translation quality control and attribute manipulation.

2.3.5. Transferability of Attribute Controllers on Pretrained Multilingual Translation Models

Liu (Liu and Niehues, 2023) explored inference-time control using gradient-based classifier guidance on a pretrained model to assess the transferability of the attribute controller across multiple languages. The experiment begins by training classifiers for various attributes on decoder activations, utilizing their predictions to adjust model activations at inference time to align with the desired attributes. The experiment used COCOA-MT (Nădejde et al., 2022) and MuST-SHE (Bentivogli et al., 2020) respectively for formality and gender controlled translation. For the translation task directions, Transfer to New Target Languages and Transfer to New Source Languages are both tested. The NLLB-200 distilled 600M model (Costa-jussà, Cross, et al., 2022) served as the pretrained model, while the OPUS-100 (Zhang et al., 2020) model was used as a Transformer-based model.

For gender-controlled experiments, this study utilized the MuST-SHE dataset, available for English-French, English-Italian, and English-Spanish. In contrast, this thesis employed the MT-GenEval test and development datasets, encompassing 8 and 9 language directions, respectively. Our multilingual experiments specifically examined English to Spanish, Portuguese, German, and Dutch translations, incorporating third-language translation pairs as prompt examples, thus covering both Latin and Germanic branches of the Indo-European language family. Furthermore, while this study centers on extending a pretrained NLLB-200 model for attribute control using classifier guidance, my thesis primarily focuses on prompting LLM models to analyze their performance in attribute-controlled translation tasks.

2.3.6. Enhanced Prompting for Attribute-Controlled Translation

Sarti (Sarti et al., 2023) proposed RAMP (Sarti et al., 2023) (Retrieval and Attribute-Marking Enhanced Prompting) method that enhances Attribute-Controlled Translation by utilizing LLMs models in few-shot and zero-shot settings. RAMP improves generation accuracy compared to standard prompting by incorporating two key components: (1) a Semantic Similarity Retrieval system that selects relevant in-context examples that will be used in a descending order in terms of similarity to the source sentence and (2) Attribute Marking which uses annotations to specify words that is related to attribute, allowing the model to better understand and apply

the desired attributes during translation. Specifically for one prompt, after the translation instruction, an extra sentence will be added that specifies the text spans that convey the desired attribute. For instance, the prompt for formality control is written as :" Given a sentence x, its translation y can be generated in a specific style. The translated sentence conveys the desired style by incorporating words such as w_1 and w_2." Here w_1 and w_2 will indicate the formality label.

The experiment used COCOA-MT (Nădejde et al., 2022) and MT-GenEval (Currey et al., 2022) respectively for formality and gender controlled translation. For LLMs, it used XGLM (Lin et al., 2022), BLOOM (Le Scao et al., 2023) and GPT-NEOX (Black et al., 2022).

Different from this study, my thesis does not assess the similarity between example translations and the source sentence, resulting in no similarity-based ordering of examples. All examples in my prompts are selected randomly. Furthermore, my prompts do not include explicit attribute marking, requiring the LLM to identify attribute-related words without any direct cues.

3. Approaches

We propose to implement three modules: the translation module, the post-editing module, and the module for detecting missing attribute information.

In the context of investigating translation and post-editing performance, comparing NLLB's encoder-decoder approach with the autoregressive decoding of LLMs will help to highlight the strengths and limitations of both architectures in producing accurate, attribute-controlled translations across various languages.

For detecting missing attribute information, we aim to assess the ability of LLMs to identify ambiguous entities that are crucial to the translation task.

3.1. Dedicated model

3.1.1. NLLB

NLLB (No Language Left Behind) is a conventional neural machine translation (NMT) model that employs the traditional encoder-decoder architecture. In this setup, encoder transforms the source token sequence into a sequence of token embeddings. The decoder attends to the encoder output and autoregressively generates the target sentence token by token. NLLB is trained on large-scale multilingual datasets and is capable of translating between 202 languages, with a focus on underrepresented ones.

3.2. LLM-based translation module

NMT models like NLLB are traditionally designed to process an input sequence and generate its translation without explicit attribute control. However, translation can be ambiguous when the target language depends on attributes such as formality or gender that are absent in the source language. Unlike standard NMT models, LLMs enable attribute-conditioned translation through prompt design, allowing users to specify desired output characteristics.

LLMs typically follow a decoder-only architecture and generate text in an autoregressive manner, meaning they produce one token at a time based on previous tokens. This differs from the encoder-decoder paradigm, where input and output sequences are processed in parallel. Despite this difference, LLMs have shown strong performance in translation tasks when guided by the prompts.

We explore two prompting strategies: zero-shot and few-shot translation. In the zero-shot setup, the model receives only task instructions and attribute specifications, without seeing any example translations. In the few-shot setup, the prompt includes a small number of example translations from a development dataset. This approach is defined as in-context learning (Brown et al., 2020), where a pre-trained LLM generalizes patterns from provided examples at inference time, without requiring fine-tuning. Previous work (Brown et al., 2020) has demonstrated that LLMs such as GPT-3, can effectively perform effective translation through few-shot learning without updating their parameters.

3.2.1. Zero-shot

The Llama prompt consists of the system prompt and the user prompt. System prompt provides initial instructions and constraints for the model, defining its behavior and role. User prompt provides the actual input from the user, specifying the request or task for the model to respond to.

In the zero-shot setup, the system prompt S includes clear instructions about the translation task, the target language l, and the desired attribute a. The source sentence x is placed in the user prompt. The LLM then generates the hypothesis sentence $h^{\text{LLM-zero-shot}}$ based on these instructions. The function $f()$, parameterized by an LLM, represents this process as:

$$h^{\text{LLM-zero-shot}} = f(S, l, a, x) \tag{3.1}$$

3.2.2. Few-shot

In the few-shot setup, we provide a sequence of k labeled translation pairs from the development or training dataset. The system prompt (S used in the zero-shot setup remains the same, and the translation pairs (x_i, y_i) are appended directly after the system prompt. Each translation pair is represented as (x_i, y_i), where x_i is the i-th source sentence, and y_i is the corresponding translation from the reference set. The language l and attribute a remain the same for all translation pairs.

For the translation task, the source sentence x_{k+1} is placed in the user prompt, and the LLM is expected to generate the hypothesis $h_{k+1}^{\text{LLM-few-shot}}$. This process can be expressed as:

$$h_{k+1}^{\text{LLM-few-shot}} = f(S, \{(x_1, y_1), \ldots, (x_k, y_k)\}, l, a, x_{k+1}) \tag{3.2}$$

3.2.3. Multi-lingual

Given the scarcity of supervised pre-trained models for low-resource languages and the even greater shortage of attribute-controlled datasets, we propose investigating the potential of using a third language as an exemplar. Through in-context learning, we aim to enable LLM to recognize patterns from this third language examples and transfer this knowledge to target translation.

We aim to investigate how the inclusion of translation examples in a third language influences the translation quality in the target language. Similar to the Translation few-shot module above, right after the system prompt S, where we indicate the target language l and the desired attribute a, we insert a sequence of k labeled translation pairs (x_i, y_i') where the target language is in the third language. Then, we place the source sentence x_{k+1} in the user prompt. The LLM is then expected to output the hypothesis h_{k+1}^{Multi}. This process can be expressed as:

$$h_{k+1}^{\text{Multi}} = f\left(S, \{(x_1, y_1'), \ldots, (x_k, y_k')\}, l, a, x_{k+1}\right) \tag{3.3}$$

3.3. Post-editing module

We aim to investigate the extent to which the post-editing module can enhance translation quality and enforce attribute control.

3.3.1. Post-editing Zero-shot

In a zero-shot setting, we give LLM an instruction in system prompt S that details the desired attribute a and target language l. The LLM's task is to refine the grammar of a given translation pair (x, h). Specifically, we ask it to improve the translation by ensuring accurate gender usage and consistent gender-related grammatical agreement. The original pair (x, h) comes from standard translation, and the LLM is expected to generate a better translation, $h^{\text{Post-edit zero-shot}}$. This process can be expressed as:

$$h^{\text{Post-edit zero-shot}} = f(S, (x, h), l, a) \tag{3.4}$$

3.3.2. Post-editing Few-shot

The post-editing few-shot approach employs a prompt S structure similar to the zero-shot setup, but extends it by including k input-output translation pairs in the form of (x_i, y_i) from training or development set. These pairs serve as demonstrations of the desired grammatical transformations, emphasizing gender accuracy and alignment. We insert the to-be-improved hypothesis pairs (x_{k+1}, h_{k+1}) in user prompt. The LLM is expected to generate an improved hypothesis $h_{k+1}^{\text{Post-edit few-shot}}$. This process can be expressed as:

$$h_{k+1}^{\text{Post-edit few-shot}} = f(S, \{(x_1, y_1), \ldots, (x_k, y_k)\}, l, a, (x_{k+1}, h_{k+1})) \tag{3.5}$$

3.4. Identify of missing attribute

To assess the LLM's ability to identify missing gender attributes, we initially used a binary "Yes"/"No" classification for determining whether additional attribute information was needed for translation. While accuracy averaged 97.62% across all language directions for ambiguous datasets, we suspect LLMs might be overly cautious, frequently defaulting to requesting additional information when faced with uncertainty.

To gain more nuanced insights into detection confidence, we switched to a 0-10 scale. In our instructions, we ask the LLM to rate whether a source sentence requires additional gender specification, focusing on human references and grammatical indicators like pronouns, gendered terms, and possessive adjectives. The source sentence x_{test} from the test set is provided in the user prompt, and the model outputs a rating r_{test} from 0 to 10.

This process can be expressed as:

$$r_{\text{test\_amb\_m}} = f(I, x_{\text{test\_amb\_m}}) \tag{3.6}$$

$$r_{\text{test\_unamb\_n}} = f(I, x_{\text{test\_unamb\_n}}) \tag{3.7}$$

$$r_{\text{dev\_amb\_i}} = f(I, x_{\text{dev\_amb\_i}}) \tag{3.8}$$

$$r_{\text{dev\_unamb\_j}} = f(I, x_{\text{dev\_unamb\_j}}) \tag{3.9}$$

$r_{\text{test\_amb\_m}}$ is the rating for the m-th ambiguous sentence from the test set.

$r_{\text{test\_unamb\_n}}$ is the rating for the n-th unambiguous sentence from the test set.

$r_{\text{dev\_amb\_i}}$ is the rating for the i-th ambiguous sentence from the development set.

$r_{\text{dev\_unamb\_j}}$ is the rating for the j-th unambiguous sentence from the development set.

For evaluation, we need to run ambiguous $x_{\text{dev\_amb\_i}}$ and unambiguous $x_{\text{dev\_unamb\_j}}$ source sentences from the development set. We then use the rating results $r_{\text{dev\_amb\_i}}$ and $r_{\text{dev\_unamb\_j}}$ to calculate the accuracy for each threshold t_{dev} from 0 to 10 per language direction. This accuracy measures how many ambiguous sentences receive scores above the threshold and how many unambiguous sentences receive scores below or equal to it. The optimal threshold t_{optimal} is the one that maximizes this accuracy.

The accuracy A_{dev} for a threshold t_{dev} can be computed using Accuracy function:

$$\begin{aligned} A_{\text{dev}} &= \text{Accuracy}(t_{\text{dev}}) \\ &= \frac{1}{N_{\text{amb}} + N_{\text{unamb}}} \left(\sum_{i=1}^{N_{\text{amb}}} \mathbb{I}(r_{\text{dev\_amb\_i}} > t_{\text{dev}}) + \sum_{j=1}^{N_{\text{unamb}}} \mathbb{I}(r_{\text{dev\_unamb\_j}} \leq t_{\text{dev}}) \right) \end{aligned} \tag{3.10}$$

where $\mathbb{I}$ is the indicator function, which is 1 if the condition is true, and 0 otherwise. N_{amb} and N_{unamb} are the total number of ambiguous and unambiguous sentences in this language direction, respectively.

The optimal threshold t_{opt} is the value of t_{dev} that maximizes the accuracy:

$$t_{\text{opt}} = \arg\max_{t_{\text{dev}}} \text{Accuracy}(t_{\text{dev}}) \tag{3.11}$$

Once the optimal threshold is determined, we apply it to the test dataset. The accuracy A_{test} is calculated by comparing the LLM's ratings against the optimal threshold t_{opt}, which gives:

$$\begin{aligned} A_{\text{test}} &= \text{TestAccuracy}(t_{\text{test}}) \\ &= \frac{1}{N_{\text{test\_amb}} + N_{\text{test\_unamb}}} \left(\sum_{m=1}^{N_{\text{test\_amb}}} \mathbb{I}(r_{\text{test\_amb\_m}} > t_{\text{opt}}) + \sum_{n=1}^{N_{\text{test\_unamb}}} \mathbb{I}(r_{\text{test\_unamb\_n}} \leq t_{\text{opt}}) \right) \end{aligned} \tag{3.12}$$

4. Experimental Setup

In our experiments, we explore attribute-controlled translation, post-editing, and attribute detection using the Llama (Dubey et al., 2024) model. For translation and post-editing, we focus on quality control and attribute control, specifically targeting formality and gender control. In the attribute detection task, our primary objective is to determine an optimal threshold and assess the accuracy rate in identifying ambiguity. These experiments provide a structured evaluation of the model's ability to control attributes and detect ambiguity effectively.

4.1. Dataset

In this work, we consider three types of tasks: translation tasks, post-editing tasks, and identifying missing attribute tasks. For translation tasks, we will conduct standard machine translation using NLLB, as well as translation with Llama in both zero-shot and few-shot settings, including multi-lingual translation. Post-editing tasks will involve refining translations using Llama in zero-shot and few-shot modes. For identifying missing attribute tasks, we will focus on detecting missing attributes in translated text using a contextual dataset. All tasks will be evaluated on CoCoA-MT (Nădejde et al., 2022) for formality-control and MT-GenEval for gender-control, with the missing attribute detection task specifically utilizing the contextual dataset from MT-GenEval (Currey et al., 2022).

4.1.1. Overview of CoCoA-MT (Contrastive Controlled MT)

CoCoA-MT covers formality-controlled translation in the conversational domain, where the source sentences lack explicit formality markers, but the translations need to include formality annotations (formal or informal).

The dataset consists of contrastive translations with phrase-level annotations of formality and grammatical gender in eight diverse language pairs: English (EN) → French (FR), German (DE), Hindi (HI), Italian (IT), Japanese (JA), Spanish (ES), Portuguese (PT) and Dutch (NL).

	CoCoA-MT Test set	**CoCoA-MT Train set**
# Directions	8	8
# Sent. per direct. per att.	600	400 (except for JA: 1000)
Avg sentence length	110	106
Domain	Topical-Chat, Telephony, and Call Center	Topical-Chat and Telephony

Table 4.1.: Formality control dataset statistics

4.1.2. Overview of MT-GenEval (Machine Translation Gender Evaluation)

MT-GenEval covers translations in two genders (female and male) from English (EN) into nine diverse and widely-spoken target languages: Arabic (AR), French (FR), German (DE), Hindi (HI), Italian (IT), Portuguese (PT), Russian (RU), Spanish (ES) and Dutch(NL).

Built from Wikipedia sources, it consists of two distinct subsets: a counterfactual subset and a contextual subset. The counterfactual subset features gender-specific sentences paired with their gender-flipped counterparts, comprising 600 test segments and 2400 development segments across multiple language pairs including English-Hindi, English-Italian, and English-Spanish. The contextual subset introduces profession-based sentences that are inherently gender-ambiguous but become disambiguated through preceding context, covering stereotypical female, male, and neutral professions.

	MT-GenEval			
	Counterfactual test dataset	**Counterfactual dev dataset**	**Contextual test dataset**	**Contextual dev dataset**
# Directions	8	8	9 (new direction: Dutch)	9
# Sent. per direct. per gender	300	1200	between 1078-1100	between 397-792
Avg sentence length	141	127	255 (context only: 143; src only: 109)	256 (context only: 143; src only: 110)
Domain	Refers to individuals of a single gender (female or male)	Same as test set	From Wikipedia related to professions	Same as test set

Table 4.2.: Grammatical gender control dataset statistics

4.2. Evaluation Metrics

In this work, for quality control, we use BLEU (Papineni et al., 2002) (Bilingual Evaluation Understudy) and COMET (Rei, Stewart, et al., 2020) (Cross-lingual Optimized Metric for Evaluation of Translation) for evaluation. For attribute control, we use M-ACC (Nădejde et al., 2022) (Matched-Accuracy), Gender Accuracy (Currey et al., 2022) and LLM as Gender Evaluator (Raunak et al., 2023) as evaluation methods. The former two methods we use their original methods without editing. For the latter we created our own prompts for evaluation.

4.2.1. Quality Control

4.2.1.1. BLEU

To evaluate the quality of translation result, BLEU calculates n-gram precision and penalizes overly short translations. We will use SacreBLEU[1](Post, 2018) to compute BLEU score.

[1]SacreBLEU signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.3

4.2.1.2. COMET

COMET (wmt22-comet-da|version:2.0.0) (Rei, Stewart, et al., 2020; Rei, C. de Souza, et al., 2022) is a machine translation evaluation metric that uses a pre-trained neural network model to predict the quality of translations. COMET is designed to correlate well with human judgments of translation quality. It uses contextual embeddings from models BERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019) and XLM- RoBERTa (Conneau, Khandelwal, et al., 2019), which captures deeper semantic information.

4.2.2. Attribute Control

4.2.2.1. M-ACC (Matched-Accuracy)

Matched-Accuracy is used to assess formality-controlled machine translation systems by quantifying how well system-generated translations align with the desired level of formality (formal or informal). It compares outputs against annotated reference translations that include formality-marking phrases, calculating the percentage of correctly classified translations. A translation is labeled as formal if it contains markers from the formal reference and none from the informal reference, and vice versa.

4.2.2.2. Gender Accuracy

Gender Accuracy is an automated evaluation method that leverages an reference set containing both correct translations and contrastive/counterfactual references that differ solely in gender-specific words. To quantify accuracy, words that are unique to the contrastive reference are identified by calculating the set difference between words in the contrastive reference and those in the correct reference (unique_con = $w_{\text{con}} \setminus w_{\text{ref}}$). This technique isolates gender-specific words since the correct and contrastive references are identical except for gender-related terms.

A translation is deemed incorrect if it contains any words from this contrastive-only set (unique_con $\cap\, w_{\text{hyp}} \neq \emptyset$), indicating that the translation system has used vocabulary specific to the incorrect gender. This straightforward metric enables automatic evaluation of gender accuracy in machine translation systems.

4.2.2.3. LLM as Gender Evaluator

The Gender Accuracy metric (from section 4.2.2.2) depends on overlap with the reference, meaning it may not be reliable when gendered words in translations use synonyms that do not appear in the reference. Previous research by (Lee et al., 2024) explored using ChatGPT-4 as an evaluator, finding a high correlation with human judgments. Based on these findings, we plan to use an LLM as the evaluator for gender-controlled translation. In our experiment, we will use Llama 3.1 as the evaluator and configure the prompt to classify the translation result as binary.

4.2.3. Multilingual translation

We explore how third-language translation examples influence target language quality. Our experiments include EN-ES with Portuguese examples, EN-PT with Spanish examples, EN-DE with Dutch examples, and EN-NL with German examples, using the Counterfactual Gender dataset. Since the CoCoA dataset lacks a Dutch direction, we run only the first two experiments for the Formality dataset.

4.2.4. Ambiguity Detection Evaluation

4.2.4.1. Arrangement on Dataset

For ambiguity detection, we need a dataset that contains both ambiguous and unambiguous texts within the same domain. However, such datasets are scarce. To address this, we use the contextual dataset from MT-GenEval, where each entry consists of two parts: a context sentence that clearly indicates the gender of the main sentence, followed by a <sep> symbol, and then the main sentence, which is gender ambiguous. This structure allows us to evaluate how well an LLM perceives ambiguity by leveraging cases where gender information is either explicitly provided or remains uncertain.

To effectively analyze ambiguity, we divide the dataset into two subsets. The first subset consists of unambiguous sentences—cases where the context provides sufficient information to resolve any gender-related uncertainty. The second subset consists of ambiguous sentences, where even with the given context, gender remains unclear. By structuring the data this way, we ensure that the LLM is tested on both clear-cut and uncertain cases.

To obtain results, we run an LLM on both subsets and have it rate each sentence on a scale from 0 to 10, where the score reflects how sure LLM thinks that additional attribute information is needed to produce a correct translation. Ideally, unambiguous sentences should receive low scores (close to 0), indicating that no extra information is required. In contrast, ambiguous sentences should receive high scores (close to 10), signifying that the LLM recognizes the need for additional details.

4.2.4.2. Evaluation

For evaluation, we first process the development dataset (dev set) to determine an optimal threshold, which serves as the decision boundary for distinguishing between ambiguous and unambiguous sentences. To find the best threshold, we iterate over possible values from 0 to 10 and, for each, calculate the classification accuracy. This accuracy is calculating how many ambiguous sentences receive scores above the threshold and how many unambiguous ones receive scores below or equal to it. The optimal threshold is the one that maximizes accuracy.

Once the best threshold is established, we apply it to the test dataset. By comparing the LLM's ratings against this threshold, we compute the final accuracy. This approach allows us to assess how well an LLM distinguishes between cases requiring additional attribute information and those that do not.

4.3. Model

In our experiments, we primarily use Llama 3.1 (Dubey et al., 2024) as our LLM. Based on the Transformer architecture, Llama 3.1 features 405 billion parameters, 126 layers, a token representation dimension of 16,384, and 128 attention heads. It is pre-trained on a 15 trillion-token multilingual corpus, a significant increase from the 1.8 trillion tokens used for Llama 2 (Touvron et al., 2023).

5. Results

In this chapter, we will evaluate the quality of zero-shot and few-shot results for attribute-controlled translation and post-editing modules, comparing them with conventional machine translation model. Additionally, we will examine findings from the ambiguity detection evaluation using the contextual dataset.

5.1. Overall Comparison

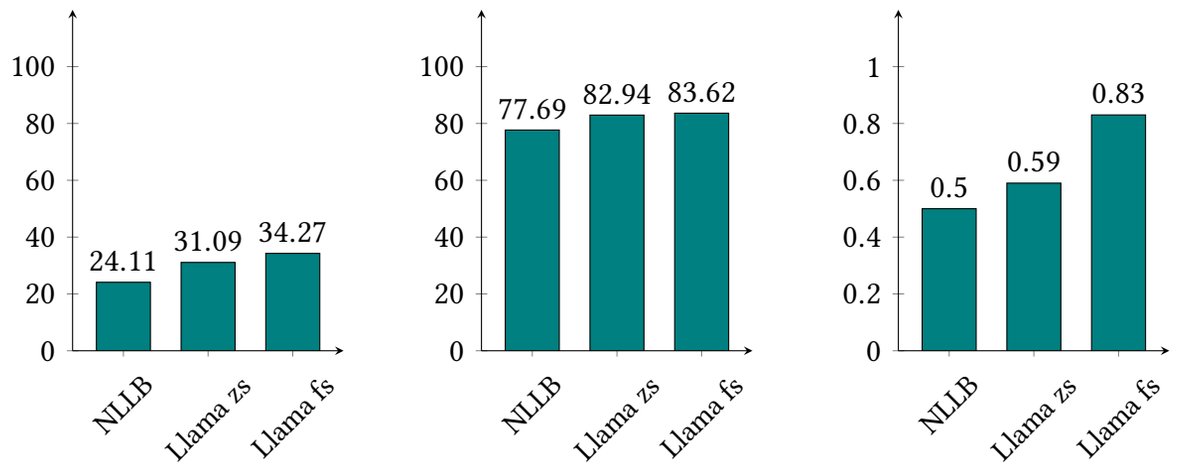
5.1.1. Formality Control

		BLEU			COMET			M-ACC		
		NLLB	Llama zs	Llama fs	NLLB	Llama zs	Llama fs	NLLB	Llama zs	Llama fs
1	DE	23.59	33.82	38.25	76.44	82.94	83.55	0.49	0.72	0.97
2	ES	31.64	36.00	39.94	81.18	84.40	84.79	0.29	0.33	0.67
3	FR	28.21	36.28	38.04	75.98	81.73	82.82	0.75	0.94	0.99
4	HI	23.39	22.49	24.49	73.15	76.62	77.46	0.92	0.91	0.97
5	IT	28.05	34.20	38.13	79.72	84.62	84.95	0.09	0.13	0.68
6	JA	7.46	21.10	22.89	75.65	83.70	84.99	0.45	0.53	0.67
7	NL	18.22	27.44	33.00	78.23	84.25	84.88	0.11	0.26	0.87
8	PT	28.58	36.98	40.11	80.03	85.14	85.41	0.48	0.98	0.96
9	Average	23.64	31.04	34.36	77.55	82.93	83.61	0.45	0.60	0.85
10	Avg_both	24.11	31.09	34.27	77.69	82.94	83.62	0.50	0.59	0.83

Table 5.1.: Comparison of NLLB and Llama in Formal dataset

		BLEU			COMET			M-ACC		
		NLLB	Llama zs	Llama fs	NLLB	Llama zs	Llama fs	NLLB	Llama zs	Llama fs
1	DE	24.16	34.25	37.96	76.35	82.73	83.40	0.51	0.69	0.90
2	ES	35.45	41.55	42.85	81.80	85.19	85.49	0.71	0.89	0.95
3	FR	24.58	31.49	34.43	75.76	81.81	82.50	0.25	0.32	0.69
4	HI	19.16	19.26	22.58	72.83	76.01	77.17	0.08	0.23	0.63
5	IT	33.70	41.34	42.06	80.86	85.54	85.93	0.91	0.96	0.98
6	JA	7.76	17.68	19.36	76.16	83.89	84.39	0.55	0.52	0.69
7	NL	23.28	32.66	34.33	78.97	84.55	85.02	0.89	0.92	0.95
8	PT	28.59	30.86	39.82	79.84	83.84	85.18	0.52	0.06	0.77
9	Average	24.58	31.14	34.17	77.82	82.95	83.64	0.55	0.57	0.82

Table 5.2.: Comparison of NLLB and Llama in Informal dataset



(a) BLEU comparison (b) COMET comparison (c) M-ACC comparison

Figure 5.1.: Comparison between NLLB, Llama zero-shot (zs), and Llama few-shot (fs) across different metrics in Formality Dataset

The comparison across different metrics in Figure 5.1 shows that the Llama few-shot setup achieves the best performance in both quality and attribute control, while NLLB performs the worst across all evaluation metrics. In terms of quality control, as seen in the BLEU comparison in Figure 5.1 (a), NLLB scores 24.11, which is 10 percentage points lower than Llama few-shot at 34.27. In the COMET comparison in Figure 5.1 (b), NLLB scores 77.69, which is 5 percentage points lower than Llama few-shot at 83.62. One possible reason for NLLB's poor quality control is the undertranslation issue, where it may skip sentences or words to avoid contradictions with the main context, resulting in a less accurate translation.

For attribute control in Figure 5.1 (c), the M-ACC score for NLLB is 0.5, which lags 33 percentage points behind Llama few-shot's 0.83, demonstrating that the few-shot setup significantly improves attribute control compared to conventional machine translation model. Additionally, Llama few-shot with M-ACC of 0.83 is 24 percentage points higher than the zero-shot setup with M-ACC of 0.59, which indicates that LLMs can learn and apply patterns to improve translations.

As can be seen in Table 5.1 and Table 5.2, all models perform relatively worse for Japanese (sixth row) and Hindi (fourth row) in terms of BLEU scores, with Japanese receiving the worst scores across all directions.

Regarding formality control, we observe that all models tend to translate French and Hindi into a formal tone, while Italian and Dutch are translated in an informal tone. As seen in the third and fourth rows of Table 5.2, both NLLB and Llama zero-shot models have low informal M-ACC scores for French and Hindi, with NLLB having the lowest informal M-ACC score of 0.08 for Hindi. The opposite pattern is seen in the formal dataset. As shown in the fifth and seventh rows of Table 5.1, both NLLB and Llama perform poorly with formal Italian and Dutch, with formal M-ACC scores below 0.26. NLLB, in particular, has the worst formal M-ACC score of 0.09 for Italian. However, we also observe that the few-shot setup significantly improves attribute control for all these languages.

We also observed that the Llama model tends to translate Portuguese in a formal tone, as seen in the eighth row of Table 5.1 with an M-ACC of 0.98 and in Table 5.2 with an M-ACC of 0.06. We will explore this language direction in more detail in Section 5.3 with the Multilingual model.

5.1.2. Grammatical Gender Control

5.1.2.1. Counterfactual gender dataset

		BLEU			COMET			LLM as Evaluator		
		NLLB	Llama zs	Llama fs	NLLB	Llama zs	Llama fs	NLLB	Llama zs	Llama fs
1	AR	25.46	17.54	19.16	82.09	78.45	79.50	0.86	0.81	0.83
2	DE	43.95	38.66	39.26	85.19	83.57	84.59	0.91	0.93	0.92
3	ES	53.60	48.77	49.65	86.32	85.02	85.65	0.79	0.88	0.87
4	FR	41.04	34.67	38.61	83.82	79.72	82.43	0.85	0.89	0.87
5	HI	30.34	21.26	21.91	78.27	72.88	73.97	0.85	0.87	0.88
6	IT	40.90	35.63	36.41	86.44	85.10	85.58	0.75	0.80	0.77
7	PT	51.44	45.44	46.13	87.93	86.79	87.15	0.86	0.93	0.89
8	RU	37.15	30.71	31.37	86.66	85.63	86.26	0.83	0.87	0.87
9	Average	40.48	34.08	35.31	84.59	82.14	83.14	0.84	0.87	0.86
10	Avg_both	41.30	34.57	35.82	84.90	82.28	83.41	0.87	0.88	0.87

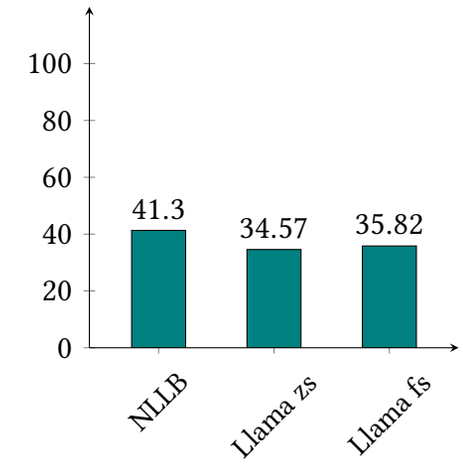
Table 5.3.: Comparison of NLLB and Llama in Counterfactual feminine dataset

		BLEU			COMET			LLM as Evaluator		
		NLLB	Llama zs	Llama fs	NLLB	Llama zs	Llama fs	NLLB	Llama zs	Llama fs
1	AR	25.74	17.85	18.77	82.37	78.46	79.50	0.87	0.81	0.81
2	DE	45.12	39.54	40.76	85.95	83.49	85.40	0.87	0.89	0.87
3	ES	54.90	49.23	50.05	87.15	85.93	86.40	0.86	0.88	0.88
4	FR	42.54	34.69	38.76	84.66	78.53	82.98	0.88	0.86	0.88
5	HI	32.60	22.18	23.11	79.37	73.32	74.45	0.93	0.91	0.90
6	IT	42.78	37.82	37.41	86.96	86.25	86.32	0.90	0.92	0.91
7	PT	53.72	47.48	49.04	88.26	87.46	87.81	0.91	0.90	0.90
8	RU	39.51	31.68	32.79	87.04	85.93	86.57	0.91	0.93	0.93
9	Average	42.11	35.06	36.34	85.22	82.42	83.68	0.89	0.89	0.89

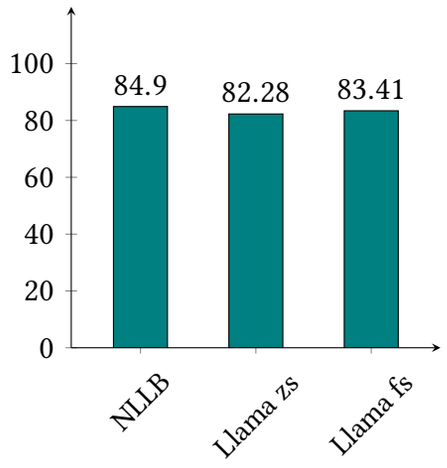
Table 5.4.: Comparison of NLLB and Llama in Counterfactual masculine dataset

		Gender Accuracy		
		NLLB	Llama zs	Llama fs
1	AR	0.71	0.84	0.83
2	DE	0.69	0.77	0.73
3	ES	0.65	0.74	0.70
4	FR	0.64	0.77	0.68
5	HI	0.58	0.73	0.71
6	IT	0.61	0.68	0.66
7	PT	0.65	0.71	0.65
8	RU	0.73	0.83	0.80
9	Average	0.66	0.76	0.72

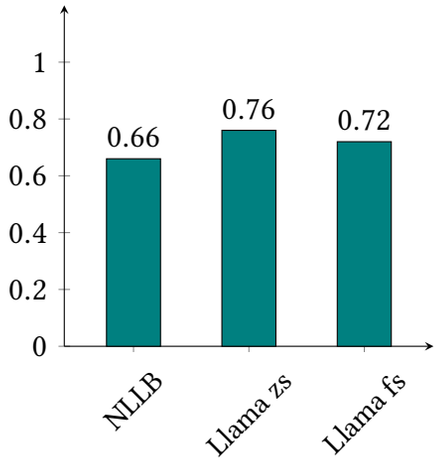
Table 5.5.: Comparison of NLLB and Llama for Gender Accuracy in Counterfactual dataset



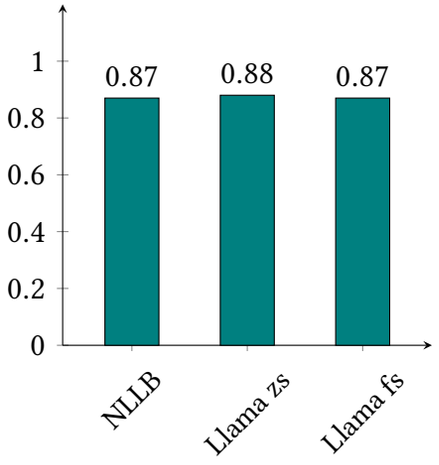
(a) BLEU comparison



(b) COMET comparison



(c) Gender Accuracy comparison



(d) LLM as Evaluator comparison

Figure 5.2.: Comparison between NLLB, Llama zero-shot (zs), and Llama few-shot (fs) across different metrics in Counterfactual Dataset

As shown in the BLEU comparison in Figure 5.2 (a) for the Counterfactual Gender dataset, NLLB performs better than the Llama models with a BLEU score of 41.3, which is 5 percentage points higher. In the COMET comparison in Figure 5.2 (b), all models perform similarly, with scores around 83.

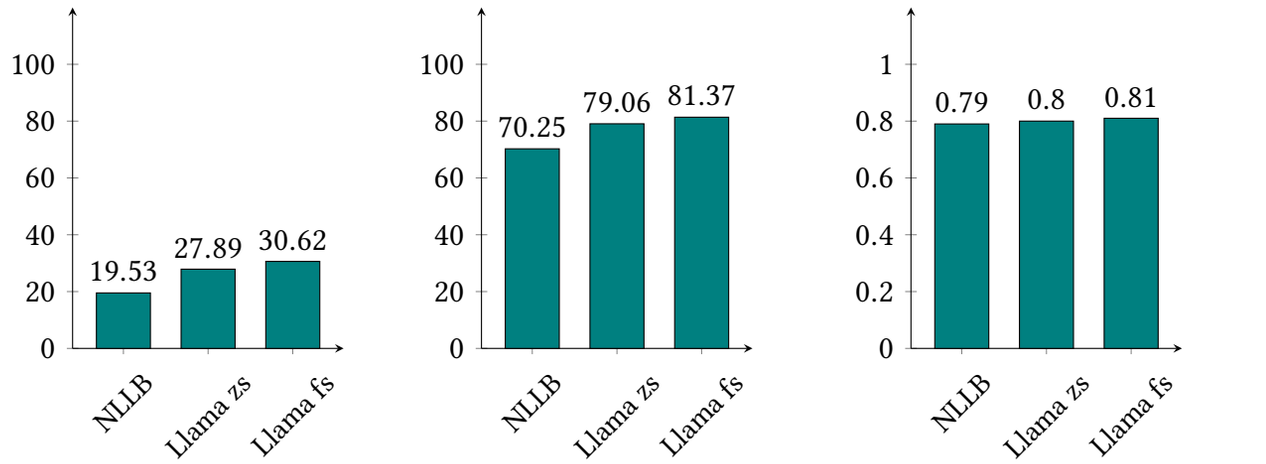
For gender control in Figure 5.2 (c), we observe two key points: Llama few-shot has a gender accuracy of 0.72, which is 4 percentage points lower than the zero-shot model (0.76), contradicting the assumption that few-shot should improve attribute control. Additionally, NLLB performs worse than both Llama models, with a gender accuracy of 0.66, which is 6 percentage points lower. However, in the LLM as Evaluator metric in Figure 5.2 (d), all models achieve comparable scores. This could be due to the word-level nature of gender accuracy metrics, which may introduce challenges related to synonymy.

Regarding translation directions, all models struggle with Arabic and Hindi in terms of BLEU scores. As seen in the first and fifth rows of Table 5.3 and Table 5.4, all models score below 33 in BLEU score.

5.1.2.2. Contextual gender dataset

		BLEU			COMET			Gender Accuracy		
		NLLB	Llama zs	Llama fs	NLLB	Llama zs	Llama fs	NLLB	Llama zs	Llama fs
1	AR	8.43	9.41	12.48	66.60	69.91	74.66	0.87	0.89	0.87
2	DE	21.92	27.95	29.12	73.56	80.92	82.73	0.79	0.83	0.82
3	ES	25.61	43.18	45.57	72.92	83.22	84.73	0.76	0.80	0.80
4	FR	17.95	36.00	38.25	64.70	81.03	82.79	0.78	0.78	0.79
5	HI	19.96	14.93	19.80	68.07	67.53	70.58	0.68	0.74	0.78
6	IT	12.91	30.48	32.31	63.16	82.07	83.86	0.81	0.75	0.77
7	NL	23.74	30.51	33.08	74.85	82.60	84.47	0.78	0.77	0.80
8	PT	20.14	35.03	38.33	68.86	83.18	84.94	0.78	0.80	0.82
9	RU	25.10	23.49	26.63	79.57	81.05	83.53	0.84	0.87	0.88
10	Average	19.53	27.89	30.62	70.25	79.06	81.37	0.79	0.80	0.81

Table 5.6.: Comparison of NLLB and Llama in Contextual dataset



(a) BLEU comparison (b) COMET comparison (c) Gender Accuracy comparison

Figure 5.3.: Comparison between NLLB, Llama zero-shot (zs), and Llama few-shot (fs) across different metrics in Contextual Dataset

As shown in the BLEU comparison in Figure 5.3 (a), for the Contextual Gender dataset, Llama outperforms NLLB in both zero-shot and few-shot setups, with scores of 27.89 and 30.62, respectively, surpassing NLLB's 19.53 by more than 8 percentage points. In the COMET metric in Figure 5.3 (b), the results are consistent, with Llama models outperforming NLLB by 8 percentage points. The few-shot setup slightly outperforms the zero-shot setup by an average of 2 percentage points.

Additionally, the few-shot setup consistently outperforms the zero-shot setup, indicating that including more translation examples in the prompt helps the LLM recognize patterns and improve its translations. In terms of gender accuracy, all models perform similarly.

Regarding translation directions, as shown in the first row of Table 5.6, all models (NLLB, Llama zero-shot, Llama few-shot) perform suboptimally with Arabic in the BLEU metric, with their worst scores in Arabic at 8.43, 9.41, and 12.48, which are less than half of their average BLEU scores (19.53, 27.89, and 30.62, respectively). However, in all directions, the few-shot setup outperforms the zero-shot setup, indicating that the additional in-context examples help improve performance.

5.2. Impact of Post-editing

5.2.1. Formality Control

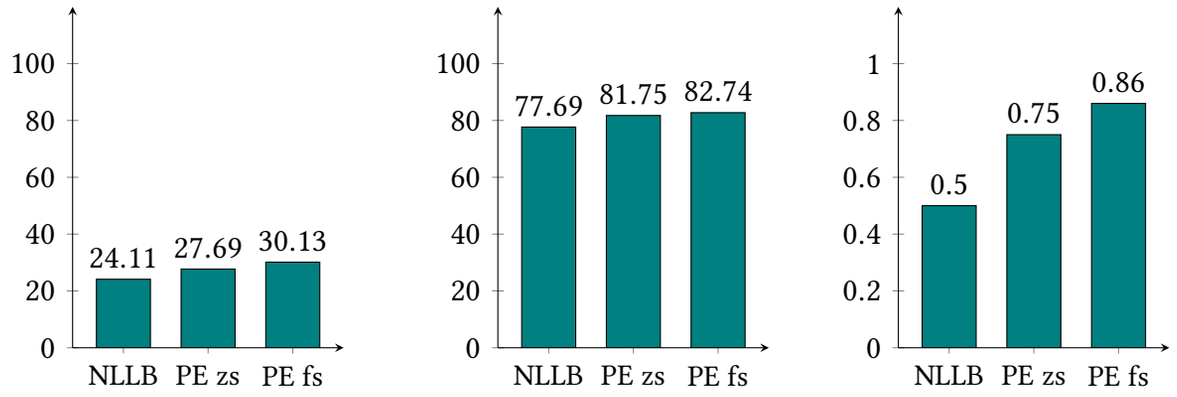
5.2.1.1. Formality dataset

		BLEU			COMET			M-ACC		
		NLLB	PE zs	PE fs	NLLB	PE zs	PE fs	NLLB	PE zs	PE fs
1	DE	23.59	29.15	31.63	76.44	81.25	82.54	0.49	0.86	0.95
2	ES	31.64	35.68	37.08	81.18	83.37	84.21	0.29	0.70	0.65
3	FR	28.21	32.89	33.78	75.98	80.41	81.54	0.75	0.98	0.95
4	HI	23.39	23.68	25.21	73.15	78.28	78.63	0.92	0.98	0.98
5	IT	28.05	31.89	35.37	79.72	83.31	84.62	0.09	0.29	0.53
6	JA	7.46	18.02	19.10	75.65	83.71	84.39	0.45	0.58	0.68
7	NL	18.22	24.56	30.07	78.23	82.94	83.86	0.11	0.54	0.95
8	PT	28.58	32.56	35.16	80.03	83.82	84.36	0.48	0.997	0.934
9	Average	23.64	28.55	30.92	77.55	82.13	83.02	0.45	0.74	0.83
10	Avg_both	24.11	27.69	30.13	77.69	81.75	82.74	0.50	0.75	0.86

Table 5.7.: Comparison of NLLB, Post-edit zero-shot (PE zs) and Post-edit few-shot (PE fs) in Formal dataset

		BLEU			COMET			M-ACC		
		NLLB	PE zs	PE fs	NLLB	PE zs	PE fs	NLLB	PE zs	PE fs
1	DE	24.16	28.21	30.21	76.35	80.72	81.41	0.51	0.89	0.94
2	ES	35.45	34.85	36.69	81.80	82.61	84.10	0.71	0.97	0.97
3	FR	24.58	28.93	30.81	75.76	79.62	81.21	0.25	0.78	0.84
4	HI	19.16	20.75	23.56	72.83	77.58	78.33	0.08	0.49	0.74
5	IT	33.70	36.69	38.02	80.86	84.12	85.00	0.91	0.99	0.98
6	JA	7.76	12.69	14.05	76.16	81.96	83.17	0.55	0.79	0.86
7	NL	23.28	25.70	27.73	78.97	82.04	83.30	0.89	0.99	0.98
8	PT	28.59	26.76	33.60	79.84	82.24	83.22	0.52	0.23	0.79
9	Average	24.58	26.82	29.33	77.82	81.36	82.47	0.55	0.77	0.89

Table 5.8.: Comparison of NLLB, Post-edit zero-shot (PE zs) and Post-edit few-shot (PE fs) in Informal dataset



(a) BLEU comparison (b) COMET comparison (c) M-ACC comparison

Figure 5.4.: Comparison between NLLB, Post-edit zero-shot (PE zs) and Post-edit few-shot (PE fs) across different metrics in Formality Dataset

Overall, as shown in Figure 5.4, Post-editing models outperform NLLB across all metrics, with the Post-editing few-shot setup delivering the best performance. In the BLEU comparison in Figure 5.4 (a), Post-editing zero-shot improves quality control by over 3 percentage points, from 24.11 to 27.69, and in the COMET comparison in Figure 5.4 (b), it shows an improvement of over 4 percentage points, from 77.69 to 81.75. For formality control, as seen in the M-ACC comparison in Figure 5.4 (c), Post-editing boosts NLLB's M-ACC score from 0.5 to 0.75 in zero-shot and 0.86 in few-shot, representing an improvement of more than 25 percentage points.

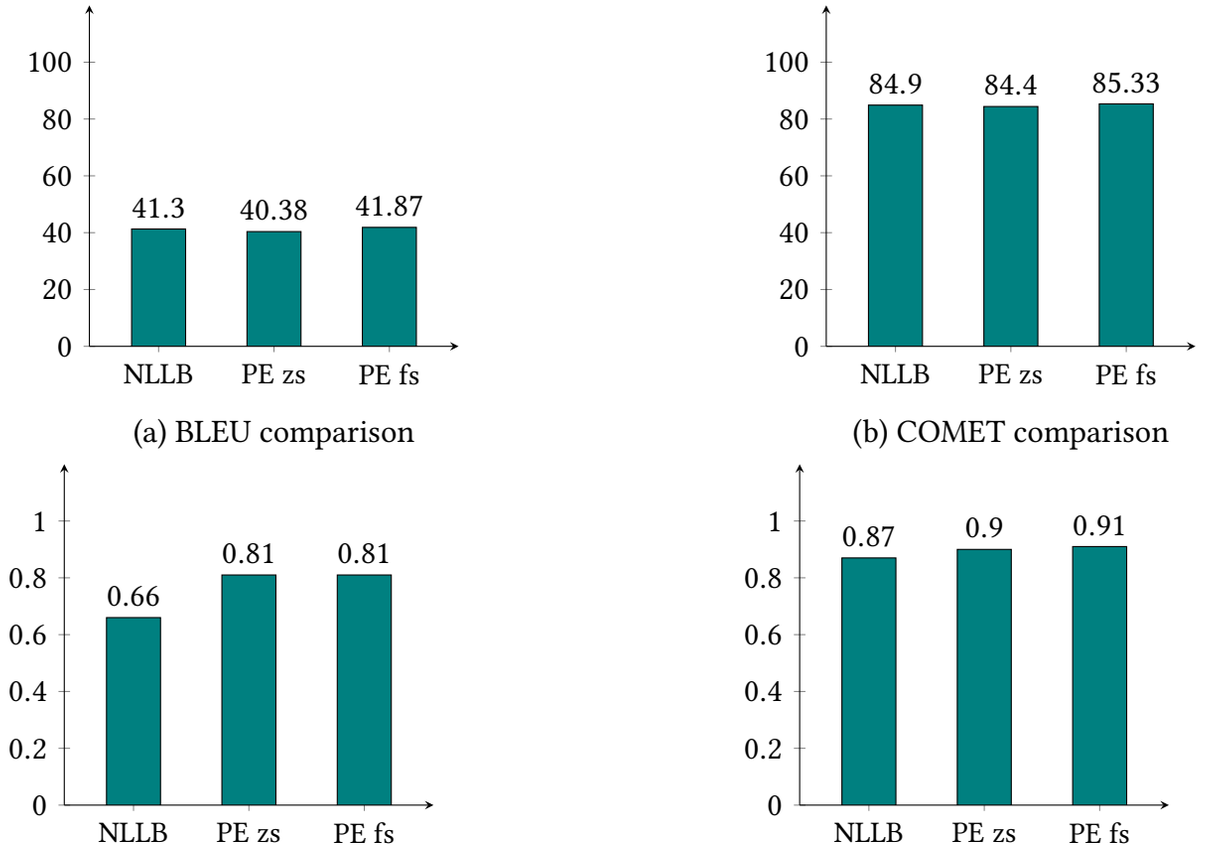
Regarding translation directions in Tables 5.7 and 5.8, both Post-editing models perform similarly to NLLB in quality control. While all models perform worse for Japanese in BLEU (sixth row of both tables), Post-editing still outperforms NLLB by over 10 percentage points in the formal dataset and 4 percentage points in the informal dataset: In the formal dataset, zero-shot and few-shot Post-editing score 18.02 and 19.10, while NLLB reaches only 7.46. In the informal dataset, Post-editing achieves 12.69 (zero-shot) and 14.05 (few-shot), compared to NLLB's 7.76.

For attribute control, Post-editing models generally improve M-ACC significantly. For example, as seen in the second, fifth, and seventh rows of Table 5.7, in formal directions, zero-shot Post-editing raises scores from 0.29 to 0.70 (Spanish), 0.09 to 0.29 (Italian), and 0.11 to 0.54 (Dutch). As seen in the fourth row of Table 5.8, in informal directions, it increases M-ACC of Hindi from 0.08 to 0.49. In these directions, Post-editing zero-shot improves M-ACC by at least 20 percentage points. Post-editing few-shot further improves the score, except in formal Spanish (second row) and formal Portuguese (eighth row), where the score drops by roughly 5 percentage points but still outperforms NLLB.

However, as observed in the eighth row of both tables, Portuguese is an outlier. The informal M-ACC score for Post-editing in informal Portuguese (in Table 5.8) is 0.23, lower than NLLB's 0.52. In contrast, the formal M-ACC score for formal Portuguese (in Table 5.7) is 0.997, close to 1. This suggests that Post-editing favors a formal tone in Portuguese.

5.2.2. Grammatical Gender Control

5.2.2.1. Counterfactual dataset



(a) BLEU comparison

(b) COMET comparison

(c) Gender Accuracy comparison

(d) LLM as Evaluator comparison

Figure 5.5.: Comparison between NLLB, Post-edit zero-shot (PE zs) and Post-edit few-shot (PE fs) across different metrics in Counterfactual Dataset

		Gender Accuracy		
		NLLB	PE zs	PE fs
1	AR	0.71	0.86	0.87
2	DE	0.69	0.84	0.81
3	ES	0.65	0.82	0.81
4	FR	0.64	0.77	0.78
5	HI	0.58	0.77	0.75
6	IT	0.61	0.75	0.76
7	PT	0.65	0.80	0.78
8	RU	0.73	0.91	0.90
9	Average	0.66	0.81	0.81

Table 5.9.: Comparison of NLLB, Post-edit zero-shot (PE zs) and Post-edit few-shot (PE fs) for Gender Accuracy in Counterfactual dataset

		BLEU			COMET			LLM as Evaluator		
		NLLB	PE zs	PE fs	NLLB	PE zs	PE fs	NLLB	PE zs	PE fs
1	AR	25.46	24.96	26.19	82.09	82.40	82.77	0.86	0.90	0.90
2	DE	43.95	44.08	45.35	85.19	85.45	85.99	0.91	0.95	0.94
3	ES	53.60	53.06	54.88	86.32	85.93	86.56	0.79	0.87	0.89
4	FR	41.04	41.23	42.34	83.82	82.62	84.31	0.85	0.88	0.90
5	HI	30.34	28.85	31.59	78.27	76.65	78.63	0.85	0.91	0.93
6	IT	40.90	39.80	42.01	86.44	85.12	86.91	0.75	0.85	0.86
7	PT	51.44	50.88	51.34	87.93	87.69	87.99	0.86	0.91	0.92
8	RU	37.15	37.13	37.64	86.66	87.12	87.39	0.83	0.90	0.91
9	Average	40.48	40.00	41.42	84.59	84.12	85.07	0.84	0.90	0.91
10	Avg_both	41.30	40.38	41.87	84.90	84.40	85.33	0.87	0.90	0.91

Table 5.10.: Comparison of NLLB, Post-edit zero-shot (PE zs) and Post-edit few-shot (PE fs) in Counterfactual feminine dataset

		BLEU			COMET			LLM as Evaluator		
		NLLB	PE zs	PE fs	NLLB	PE zs	PE fs	NLLB	PE zs	PE fs
1	AR	25.74	24.66	25.84	82.37	82.63	83.26	0.87	0.89	0.89
2	DE	45.12	44.17	45.75	85.95	85.81	86.43	0.87	0.90	0.89
3	ES	54.90	53.74	55.42	87.15	86.26	87.34	0.86	0.88	0.89
4	FR	42.54	42.23	42.85	84.66	83.47	84.87	0.88	0.88	0.90
5	HI	32.60	30.06	32.35	79.37	77.54	79.26	0.93	0.95	0.94
6	IT	42.78	41.07	43.04	86.96	86.38	87.26	0.90	0.91	0.93
7	PT	53.72	52.10	53.28	88.26	87.78	88.32	0.91	0.90	0.91
8	RU	39.51	38.13	40.08	87.04	87.57	87.92	0.91	0.92	0.94
9	Average	42.11	40.77	42.32	85.22	84.68	85.58	0.89	0.90	0.91

Table 5.11.: Comparison of NLLB, Post-edit zero-shot (PE zs) and Post-edit few-shot (PE fs) in Counterfactual masculine dataset

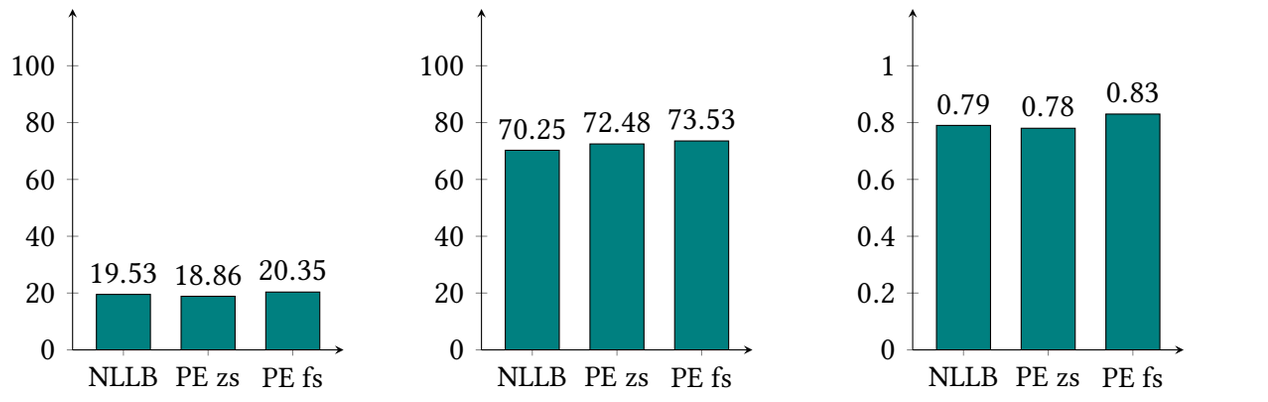
As shown in BLEU and COMET metrics in Figure 5.5 (a) and (b), NLLB and Post-editing models perform similarly in terms of quality control, as Post-editing maintains the quality of the original NLLB translation.

For attribute control in Figure 5.5 (c) and (d), Post-editing outperforms NLLB, with an average improvement of over 15 percentage points in gender accuracy and 3 percentage points in the LLM as Evaluator metric. This demonstrates that Post-editing is highly effective in enhancing attribute control.

5.2.2.2. Contextual dataset

		BLEU			COMET			LLM as Evaluator		
		NLLB	PE zs	PE fs	NLLB	PE zs	PE fs	NLLB	PE zs	PE fs
1	AR	8.43	7.74	7.96	66.60	69.02	68.73	0.87	0.84	0.90
2	DE	21.92	20.75	22.60	73.56	75.96	76.86	0.79	0.71	0.80
3	ES	25.61	24.47	27.89	72.92	72.46	74.60	0.76	0.71	0.78
4	FR	17.95	18.15	20.93	64.70	67.86	70.20	0.78	0.80	0.83
5	HI	19.96	18.83	17.79	68.07	68.42	67.38	0.68	0.77	0.79
6	IT	12.91	12.62	14.90	63.16	66.98	69.46	0.81	0.82	0.83
7	NL	23.74	23.06	25.24	74.85	77.85	78.96	0.78	0.75	0.81
8	PT	20.14	19.11	20.60	68.86	72.09	73.35	0.78	0.84	0.79
9	RU	25.10	25.03	25.22	79.57	81.66	82.22	0.84	0.79	0.90
10	Average	19.53	18.86	20.35	70.25	72.48	73.53	0.79	0.78	0.83

Table 5.12.: Comparison of NLLB, Post-edit zero-shot (PE zs) and Post-edit few-shot (PE fs) in Contextual dataset



(a) BLEU comparison (b) COMET comparison (c) LLM as Evaluator comparison

Figure 5.6.: Comparison between NLLB, Post-edit zero-shot (PE zs) and Post-edit few-shot (PE fs) across different metrics in Contextual Dataset

For the Contextual dataset in Figure 5.6 (a) and (b), Post-editing performs similarly to or slightly better than NLLB, as the LLM is instructed to maintain translation quality. In BLEU all models perform around 19.6. In COMET metric Post-editing zero-shot and few-shot improve NLLB's score by 2 and 3 percentage points, raising it from 70.25 to 72.48 and 73.53, respectively.

For attribute control in Figure 5.6 (c), NLLB and Llama zero-shot achieve comparable scores of 0.79 and 0.78, respectively. Post-editing few-shot further improves NLLB's translation by 4 percentage points, increasing from 0.79 to 0.83.

Regarding translation directions in Table 5.12, Post-editing closely mirrors NLLB in BLEU since it refines the given translation.

5.3. Impact of multilingual prompts

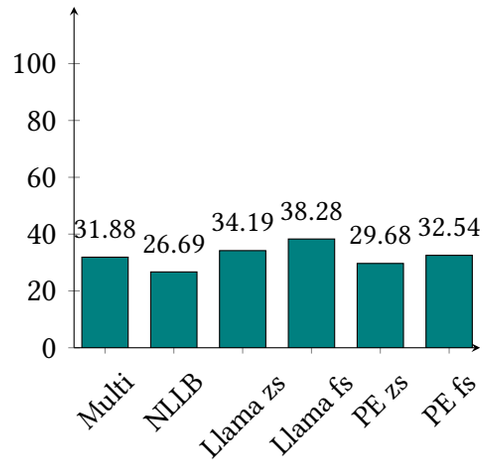
5.3.1. Formality dataset

			Directions					
		Model	EN→DE	EN→NL	EN→ES	EN→PT	Average	Avg_Both
1	BLEU	Multi-ling	32.45	24.68	37.05	31.39	31.39	31.88
2		NLLB	23.59	18.22	31.64	28.58	25.51	26.69
3		Llama zs	33.82	27.44	36.00	36.98	33.56	34.19
4		Llama fs	38.25	33.00	39.94	40.11	37.83	38.28
5		PE zs	29.15	24.56	35.68	32.56	30.49	29.68
6		PE fs	31.63	30.07	37.08	35.16	33.49	32.54
7	COMET	Multi-ling	82.08	81.86	84.47	83.36	82.94	83.03
8		NLLB	76.44	78.23	81.18	80.03	78.97	79.11
9		Llama zs	82.94	84.25	84.40	85.14	84.18	84.13
10		Llama fs	83.55	84.88	84.79	85.41	84.66	84.72
11		PE zs	81.25	82.94	83.37	83.82	82.85	82.37
12		PE fs	82.54	83.86	84.21	84.36	83.74	83.38
13	M-ACC	Multi-ling	0.96	0.74	0.65	0.99	0.84	0.79
14		NLLB	0.49	0.11	0.29	0.48	0.34	0.50
15		Llama zs	0.72	0.26	0.33	0.98	0.57	0.61
16		Llama fs	0.97	0.87	0.67	0.96	0.87	0.88
17		PE zs	0.86	0.54	0.70	1.00	0.77	0.77
18		PE fs	0.95	0.95	0.65	0.93	0.87	0.90

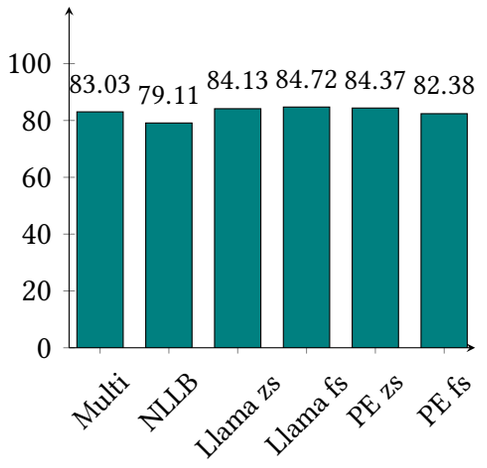
Table 5.13.: Comparison of Multi-lingual with all models in formal dataset

			Directions				
		Model	EN→DE	EN→NL	EN→ES	EN→PT	Average
1	BLEU	Multi-ling	30.34	32.35	39.01	27.76	32.36
2		NLLB	24.16	23.28	35.45	28.59	27.87
3		Llama zs	34.25	32.66	41.55	30.86	34.83
4		Llama fs	37.96	34.33	42.85	39.82	38.74
5		PE zs	28.21	25.70	34.85	26.76	28.88
6		PE fs	30.21	27.73	34.85	33.60	31.60
7	COMET	Multi-ling	80.92	84.27	84.39	82.86	83.11
8		NLLB	76.35	78.97	81.80	79.84	79.24
9		Llama zs	82.73	84.55	85.19	83.84	84.08
10		Llama fs	83.40	85.02	85.49	85.18	84.77
11		PE zs	80.72	82.04	82.61	82.24	81.90
12		PE fs	81.41	83.30	84.10	83.22	83.01
13	M-ACC	Multi-ling	0.95	0.98	0.98	0.09	0.75
14		NLLB	0.51	0.89	0.71	0.52	0.66
15		Llama zs	0.69	0.92	0.89	0.06	0.64
16		Llama fs	0.90	0.95	0.95	0.77	0.89
17		PE zs	0.89	0.99	0.97	0.23	0.77
18		PE fs	0.94	0.98	0.97	0.79	0.92

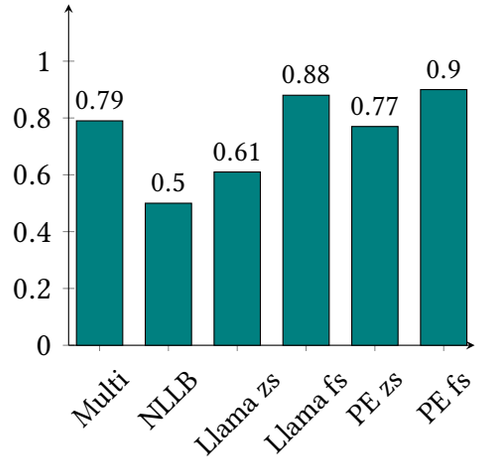
Table 5.14.: Comparison of Multi-lingual with all models in informal dataset



(a) BLEU comparison



(b) COMET comparison



(c) M-ACC comparison

Figure 5.7.: Comparison between Multilingual, NLLB, Llama zero-shot (Llama zs), Llama few-shot (Llama fs), Post-edit zero-shot (PE zs), and Post-edit few-shot (PE fs) across different metrics in Formality Dataset

For quality control in the CoCoA formality dataset, Llama models perform best overall as seen in Figure 5.7 (a) and (b). The Multilingual model surpasses NLLB, with BLEU scores increasing from 26.69 to 31.88 and COMET scores rising from 79.11 to 83.03.

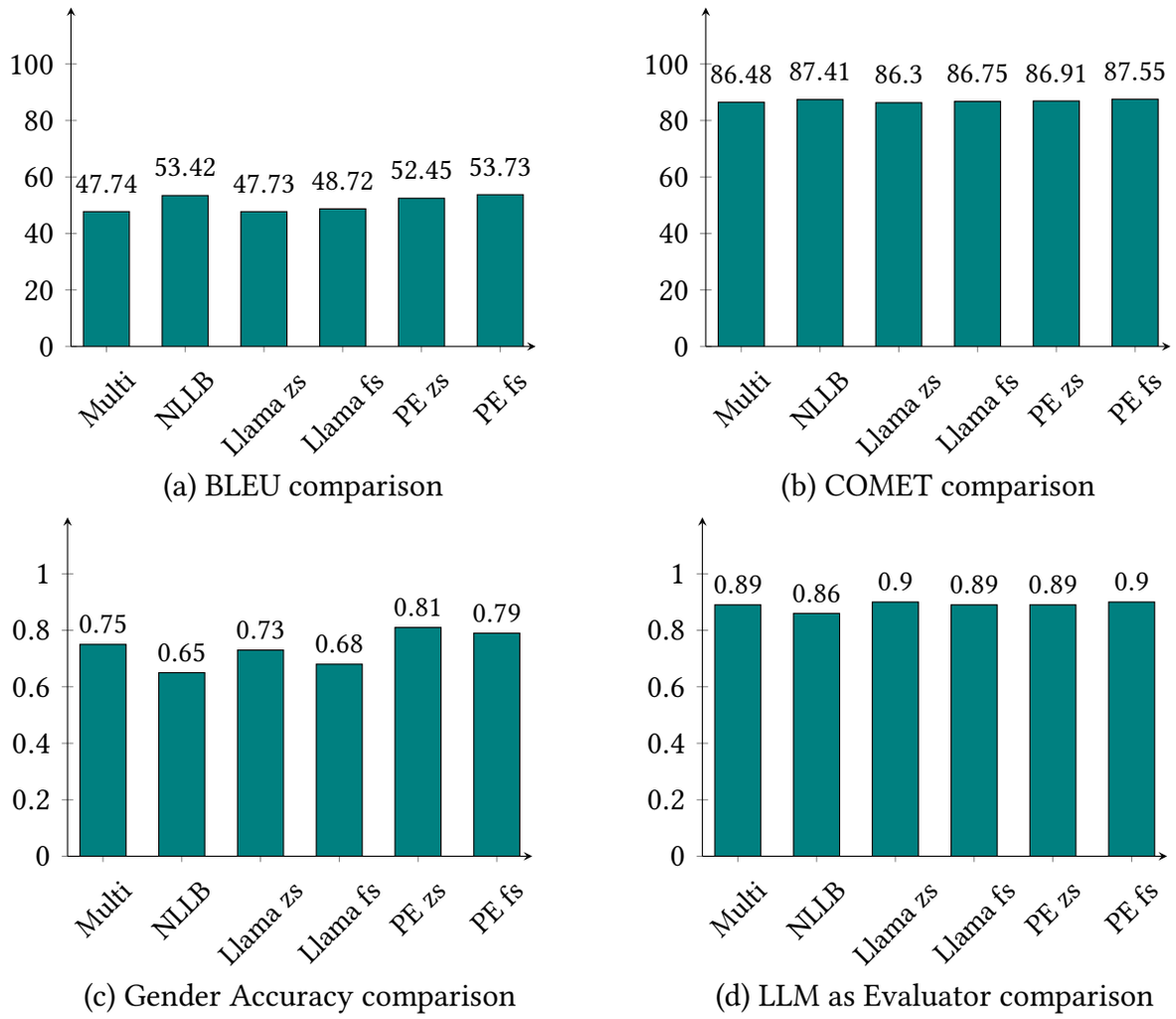
For attribute control in Figure 5.7 (c), the Multilingual model achieves a M-ACC of 0.79, significantly outperforming NLLB (0.5) and Llama zero-shot (0.61) by over 18 percentage points. However, Llama few-shot (0.88) and Post-editing few-shot (0.9) perform best, exceeding the Multilingual model by more than 9 percentage points.

In specific translation directions, Portuguese remains an outlier. As shown in the thirteenth row in Table 5.14, the Multilingual model scores low M-ACC of 0.09 for informal Portuguese but achieves a high M-ACC of 0.99 for formal Portuguese in thirteenth row in Table 5.13. Meanwhile, Llama few-shot (eighteenth row) performs well for informal Portuguese at 0.77. These results suggest that third-language examples are less effective than same-language examples for formality control.

5.3.2. Counterfactual gender dataset

For quality control, the Multilingual model underperforms compared to NLLB, with BLEU dropping 5 percentage points from 53.42 to 47.74 in Figure 5.8 (a). However, COMET scores in Figure 5.8 (b) show comparative performance across all models.

For gender control in Figure 5.8 (c), the Multilingual model achieves 0.75 accuracy, surpassing all models except Post-editing (0.81 for zero-shot, 0.79 for few-shot). This suggests that third-language examples help the LLM learn patterns and improve gender control. But Post-editing will be more effective in gender control.



(a) BLEU comparison

(b) COMET comparison

(c) Gender Accuracy comparison

(d) LLM as Evaluator comparison

Figure 5.8.: Comparison between Multilingual, NLLB, Llama zero-shot (Llama zs), Llama few-shot (Llama fs), Post-edit zero-shot (PE zs), and Post-edit few-shot (PE fs) across different metrics in Counterfactual Dataset

		Model	Feminine source			Masculine source			Avg_Both
			EN→ES	EN→PT	Average	EN→ES	EN→PT	Average	
1	BLEU	Multi-ling	49.01	45.01	47.01	50.05	46.90	48.48	47.74
2		NLLB	53.60	51.44	52.52	54.90	53.72	54.31	53.42
3		Llama zs	48.77	45.44	47.10	49.23	47.48	48.36	47.73
4		Llama fs	49.65	46.13	47.89	50.05	49.04	49.55	48.72
5		PE zs	53.06	50.88	51.97	53.74	52.10	52.92	52.45
6		PE fs	54.88	51.34	53.11	55.42	53.28	54.35	53.73
7	COMET	Multi-ling	85.33	86.76	86.04	86.46	87.37	86.91	86.48
8		NLLB	86.32	87.93	87.12	87.15	88.26	87.71	87.41
9		Llama zs	85.02	86.79	85.90	85.93	87.46	86.69	86.30
10		Llama fs	85.65	87.15	86.40	86.40	87.81	87.11	86.75
11		PE zs	85.93	87.69	86.81	86.26	87.78	87.02	86.91
12		PE fs	86.56	87.99	87.28	87.34	88.32	87.83	87.55
13	Gender Accuracy	Multi-ling	0.77	0.73	0.75	-	-	-	-
14		NLLB	0.65	0.65	0.65	-	-	-	-
15		Llama zs	0.74	0.71	0.73	-	-	-	-
16		Llama fs	0.70	0.65	0.68	-	-	-	-
17		PE zs	0.82	0.80	0.81	-	-	-	-
18		PE fs	0.81	0.78	0.79	-	-	-	-
19	LLM as Evaluator	Multi-ling	0.89	0.92	0.91	0.87	0.88	0.88	0.89
20		NLLB	0.79	0.86	0.83	0.86	0.91	0.88	0.86
21		Llama zs	0.88	0.93	0.91	0.88	0.90	0.89	0.90
22		Llama fs	0.87	0.89	0.88	0.88	0.90	0.89	0.89
23		PE zs	0.87	0.91	0.89	0.88	0.90	0.89	0.89
24		PE fs	0.89	0.92	0.90	0.89	0.91	0.90	0.90

Table 5.15.: Comparison of Multi-lingual with all models in Counterfactual dataset

5.3.3. Target Language Error

In our experiments with the formality and gender datasets, we observed that outputs often contained multiple languages. To address this, we used langid (Lui and Baldwin, 2012) to measure target language accuracy. As shown in tenth row in Table 5.16 and Table 5.18, the Multilingual model had the lowest accuracy, with 90.31% for formality and 95.67% for gender, while other models exceeded 98% in both datasets. This suggests that incorrect target language outputs may contribute to the Multilingual model's suboptimal performance.

We also analyzed the incorrect target languages. As seen in Table 5.20, the third-language examples sometimes mixed with the target language, potentially affecting quality. For example, Dutch appeared in EN-DE outputs in the third row, and German in EN-NL outputs in the fourth row. This suggests a possible trade-off between translation quality and gender control in the Multilingual model.

		Target Language Accuracy					
		NLLB	Llama zs	Llama fs	Multi-ling	PE zs	PE fs
1	DE	99.33%	99.50%	99.33%	92.17%	99.33%	99.17%
2	ES	97.00%	97.33%	97.50%	94.67%	97.00%	96.83%
3	FR	99.33%	98.67%	99.83%	-	99.83%	99.83%
4	HI	98.67%	97.83%	99.00%	-	98.83%	98.50%
5	IT	97.67%	99.00%	99.33%	-	99.17%	99.00%
6	JA	100.00%	100.00%	100.00%	-	100.00%	100.00%
7	NL	96.31%	98.99%	99.50%	79.23%	98.66%	98.49%
8	PT	94.82%	98.33%	97.50%	90.48%	98.00%	96.66%
9	Average	97.89%	98.71%	99.00%	89.14%	98.85%	98.56%
10	Avg_both	97.89%	98.74%	98.75%	90.31%	98.41%	98.06%

Table 5.16.: Comparison of target language accuracy in formal dataset

		Target Language Accuracy					
		NLLB	Llama zs	Llama fs	Multi-ling	PE zs	PE fs
1	DE	99.33%	99.50%	99.67%	85.83%	99.00%	98.83%
2	ES	97.00%	97.00%	96.67%	92.83%	95.67%	95.50%
3	FR	99.33%	99.33%	99.67%	-	99.33%	99.00%
4	HI	98.67%	97.67%	97.50%	-	97.50%	96.17%
5	IT	97.67%	99.33%	99.33%	-	98.83%	98.83%
6	JA	100.00%	100.00%	100.00%	-	100.00%	100.00%
7	NL	96.31%	98.99%	98.83%	96.31%	97.32%	97.65%
8	PT	94.82%	98.33%	96.33%	90.98%	96.16%	94.49%
9	Average	97.89%	98.77%	98.50%	91.49%	97.98%	97.56%

Table 5.17.: Comparison of target language accuracy in informal dataset

		Target Language Accuracy					
		NLLB	Llama zs	Llama fs	Multi-ling	PE zs	PE fs
1	AR	100.00%	98.33%	98.33%	-	99.67%	99.67%
2	DE	99.67%	98.67%	98.67%	-	99.67%	99.67%
3	ES	98.00%	97.67%	97.67%	96.33%	98.67%	98.33%
4	FR	99.33%	94.00%	98.67%	-	100.00%	100.00%
5	HI	97.33%	98.00%	98.67%	-	97.33%	97.67%
6	IT	99.67%	100.00%	99.67%	-	99.00%	100.00%
7	PT	97.33%	97.00%	96.67%	94.00%	99.00%	98.00%
8	RU	97.67%	97.33%	97.33%	-	97.67%	97.00%
9	Average	98.63%	97.63%	98.21%	95.17%	98.88%	98.79%
10	Avg_both	98.67%	97.67%	98.54%	95.67%	99.00%	98.94%

Table 5.18.: Comparison of target language accuracy in feminine dataset

		Target Language Accuracy					
		NLLB	Llama zs	Llama fs	Multi-ling	PE zs	PE fs
1	AR	99.00%	98.00%	98.00%	-	99.67%	99.33%
2	DE	99.33%	98.67%	99.00%	-	99.67%	99.67%
3	ES	98.33%	98.00%	98.33%	96.67%	98.67%	98.67%
4	FR	100.00%	92.67%	98.67%	-	100.00%	100.00%
5	HI	97.67%	98.33%	98.33%	-	98.33%	98.67%
6	IT	99.67%	100.00%	100.00%	-	99.33%	99.67%
7	PT	97.67%	97.67%	100.00%	95.67%	98.67%	98.00%
8	RU	98.00%	98.33%	98.67%	-	98.67%	98.67%
9	Average	98.71%	97.71%	98.88%	96.17%	99.13%	99.09%

Table 5.19.: Comparison of target language accuracy in masculine dataset

		Wrong target language detected	
		Formality Dataset	Gender Dataset
1	EN->ES	GL, PT	PT,CA,GL
2	EN->PT	GL,ES,AF	ES
3	EN->DE	NL	-
4	EN->NL	DE	-

Table 5.20.: List of wrong target language found in Multilingual module

5.4. Identifying missing attribute

We explored using an LLM for binary classification to identify missing attributes in an ambiguous dataset. As shown in Table 5.21, the average accuracy in the tenth row is high at 97.62%. This suggests the LLM may be conservative in confirming sufficient attribute information for translation.

	Direction	Binary result accuracy
1	EN->AR	99.36%
2	EN->DE	99.36%
3	EN->ES	98.27%
4	EN->FR	92.45%
5	EN->HI	99.91%
6	EN->IT	97.17%
7	EN->NL	93.88%
8	EN->PT	99.36%
9	EN->RU	98.82%
10	Average	97.62%

Table 5.21.: Binary result accuracy from LLM request

	Threshold	Accuracy								
		AR	DE	ES	FR	HI	IT	NL	PT	RU
1	0	38.60%	37.99%	38.84%	38.93%	38.59%	37.35%	41.61%	38.36%	37.34%
2	1	38.60%	37.99%	38.84%	38.93%	38.59%	37.35%	41.61%	38.36%	37.34%
3	2	45.29%	43.85%	45.12%	46.17%	45.90%	44.35%	46.61%	44.29%	43.95%
4	3	45.29%	43.85%	45.12%	46.17%	45.90%	44.35%	46.61%	44.29%	43.95%
5	4	45.29%	45.44%	46.06%	46.07%	47.04%	44.99%	45.05%	44.81%	44.89%
6	5	45.29%	45.44%	46.06%	46.07%	47.04%	44.99%	45.05%	44.81%	44.89%
7	6	45.80%	45.86%	46.25%	46.27%	47.04%	45.53%	45.46%	45.25%	45.32%
8	7	45.80%	45.86%	46.25%	46.27%	47.04%	45.53%	45.46%	45.25%	45.32%
9	8	50.03%	50.13%	50.00%	50.05%	50.06%	50.05%	50.05%	49.96%	49.96%
10	9	50.03%	50.04%	50.09%	50.05%	50.06%	50.05%	50.05%	50.04%	50.04%
11	10	50.03%	50.04%	50.09%	50.05%	50.06%	50.05%	50.05%	50.04%	50.04%

Table 5.22.: Threshold and Accuracy Result of Contextual Development dataset

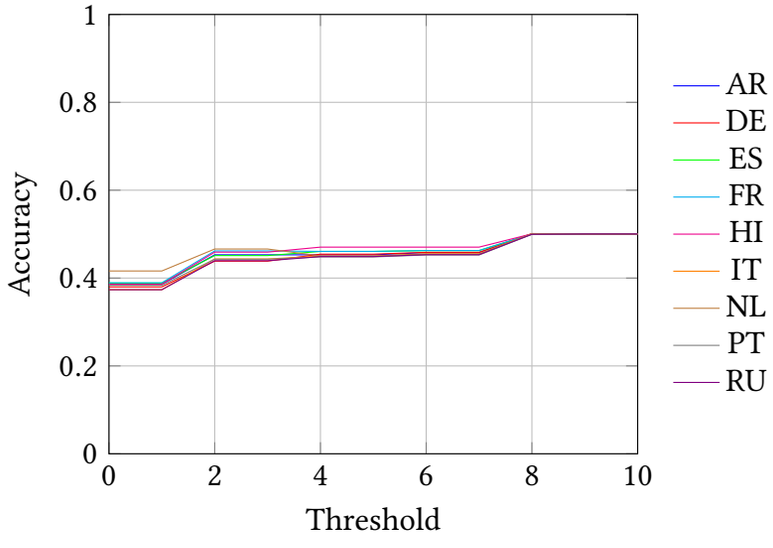


Figure 5.9.: Accuracy vs. Threshold for all directions

	Direction	Threshold	Accuracy	
			Dev dataset	Test dataset
1	EN->AR	8	50.03%	50.07%
2	EN->DE	8	50.13%	50.00%
3	EN->ES	9	50.09%	50.02%
4	EN->FR	8	50.05%	50.02%
5	EN->HI	8	50.06%	50.02%
6	EN->IT	8	50.05%	50.05%
7	EN->NL	8	50.05%	50.02%
8	EN->PT	9	50.04%	50.02%
9	EN->RU	9	50.04%	50.00%
10	Average	8.3	50.06%	50.02%

Table 5.23.: Threshold and Accuracy Result of Development and Test dataset

To further investigate its confidence in detecting missing attributes, we used a 0-10 rating scale. As seen in the ninth and tenth rows of Table 5.22, accuracy is highest at thresholds of 8 or 9 across all directions, indicating the LLM tends to request more attribute information.

However, as shown in Table 5.23 and Figure 5.9, average accuracy of test dataset in the tenth row remains around 50% across all thresholds and directions, suggesting the LLM is ineffective at identifying missing attributes.

6. Conclusion

After presenting and analyzing the results, we will address the research questions (Section 1.2) and outline directions for future work.

6.1. Answers to Research Questions

Research Question 1: How well can current state-of-the-art LLMs (e.g., Llama 3.1) achieve attribute-controlled translation into diverse target languages? How much does the performance differ under zero-shot and few-shot setups?

Llama 3.1 performs well in attribute-controlled translation across diverse languages, outperforming standard translation for formality and contextual datasets. The exception is the Counterfactual Gender dataset, where NLLB achieves higher quality scores.

Few-shot approaches generally improve both quality and attribute control compared to zero-shot, suggesting LLMs can effectively learn and apply translation patterns. It is noticed that zero-shot scores higher in gender accuracy for the counterfactual gender dataset, though this may be due to synonym issues.

However, LLMs face certain drawbacks: longer inference times due to complex architectures and larger model sizes requiring more memory and hardware resources. These limitations reduce their suitability for real-time translation applications where efficiency is crucial.

Research Question 2: To what extent can LLM-based post-editing improve the output of standard translation models (in both quality and attribute control)?

LLM-based post-editing significantly improves attribute control, particularly in the post-editing few-shot setup, which achieves the best attribute control performance. The quality control remains consistent with the standard translation, as the prompt instructs the LLM to preserve the original translation quality.

However, LLM-based post-editing has notable limitations. It requires longer inference times compared to conventional MT due to its complex architecture and larger model size. The process is further slowed by the need to pass additional parameters, such as the translation pair to be improved, making it less efficient for time-sensitive applications.

Research Question 3: To what extent can attribute-controlled translation examples from different languages help?

The Multilingual model improves attribute control but at the cost of quality control, indicating a trade-off. This decline in quality may stem from the inclusion of third-language examples, which can lead to incorrect target language outputs. When compared to Llama few-shot, same-language few-shot examples prove to be more effective for both quality and attribute control.

Research Question 4: How can we detect from the input sentence alone whether the model needs additional attribute information?

We explored binary classification and 0-10 scale rating methods, but the results were inconclusive. The LLM displayed a conservative bias in the binary classification, often requesting more attribute information. In the scale rating approach, the LLM's accuracy was inadequate, suggesting that it struggles to effectively assess the need for additional attribute information.

6.2. Future Work

For future work, investigating TOWER D. M. Alves et al. (2024) as an alternative LLM model to Llama could offer valuable comparative insights and help evaluate its effectiveness in attribute-controlled translation. Additionally, examining NLLB's undertranslation issue by processing translations on a sentence-by-sentence basis may provide a clearer understanding of this challenge and potential strategies for mitigating it.

Bibliography

Achiam, Josh et al. (2023). “Gpt-4 technical report”. In: *arXiv preprint arXiv:2303.08774* (cit. on p. 11).

Alves, Duarte M et al. (2024). “Tower: An open multilingual large language model for translation-related tasks”. In: *arXiv preprint arXiv:2402.17733* (cit. on pp. 7, 42).

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). “Neural Machine Translation by Jointly Learning to Align and Translate”. In: *CoRR* abs/1409.0473. URL: `https://api.semanticscholar.org/CorpusID:11212020` (cit. on p. 6).

Bengio, Yoshua et al. (2003). “A neural probabilistic language model”. In: *Journal of machine learning research* 3.Feb, pp. 1137–1155 (cit. on p. 3).

Bentivogli, Luisa et al. (2020). “Gender in danger? evaluating speech translation technology on the MuST-SHE corpus”. In: *arXiv preprint arXiv:2006.05754* (cit. on pp. 10, 12).

Black, Sid et al. (2022). “Gpt-neox-20b: An open-source autoregressive language model”. In: *arXiv preprint arXiv:2204.06745* (cit. on p. 13).

Brown, Tom et al. (2020). “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33, pp. 1877–1901 (cit. on p. 15).

Conneau, Alexis, Kartikay Khandelwal, et al. (2019). “Unsupervised cross-lingual representation learning at scale”. In: *arXiv preprint arXiv:1911.02116* (cit. on p. 21).

Conneau, Alexis and Guillaume Lample (2019). “Cross-lingual language model pretraining”. In: *Advances in neural information processing systems* 32 (cit. on p. 21).

Costa-jussà, Marta R, Pierre Andrews, et al. (2023). “Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale”. In: *arXiv preprint arXiv:2305.13198* (cit. on p. 11).

Costa-jussà, Marta R, James Cross, et al. (2022). “No language left behind: Scaling human-centered machine translation”. In: *arXiv preprint arXiv:2207.04672* (cit. on pp. 6, 7, 12).

Currey, Anna et al. (2022). “MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation”. In: *arXiv preprint arXiv:2211.01355* (cit. on pp. 10, 13, 19, 20).

Devlin, Jacob (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (cit. on p. 5).

Devlin, Jacob et al. (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/ (cit. on p. 21).

Dubey, Abhimanyu et al. (2024). "The llama 3 herd of models". In: *arXiv preprint arXiv:2407.21783* (cit. on pp. 6, 19, 22).

Fan, Angela et al. (2020). "Beyond English-Centric Multilingual Machine Translation". In: *J. Mach. Learn. Res.* 22, 107:1–107:48. URL: https://api.semanticscholar.org/CorpusID:224814118 (cit. on p. 11).

Freitag, Markus et al. (2021). "Experts, errors, and context: A large-scale study of human evaluation for machine translation". In: *Transactions of the Association for Computational Linguistics* 9, pp. 1460–1474 (cit. on p. 11).

Garg, Sarthak et al. (2024). "Generating Gender Alternatives in Machine Translation". In: *arXiv preprint arXiv:2407.20438* (cit. on p. 10).

Goyal, Naman et al. (2022). "The flores-101 evaluation benchmark for low-resource and multilingual machine translation". In: *Transactions of the Association for Computational Linguistics* 10, pp. 522–538 (cit. on p. 7).

Heffernan, Kevin, Onur Çelebi, and Holger Schwenk (2022). "Bitext mining using distilled sentence representations for low-resource languages". In: *arXiv preprint arXiv:2205.12654* (cit. on p. 7).

Joshi, Pratik et al. (2020). "The state and fate of linguistic diversity and inclusion in the NLP world". In: *arXiv preprint arXiv:2004.09095* (cit. on p. 7).

Kenny, Dorothy (2018). "Machine translation". In: *The Routledge handbook of translation and philosophy*. Routledge, pp. 428–445 (cit. on p. 6).

Kocmi, Tom et al. (2022). "Findings of the 2022 conference on machine translation (WMT22)". In: *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 1–45 (cit. on p. 11).

Koehn, Philipp (Sept. 2005). "Europarl: A Parallel Corpus for Statistical Machine Translation". In: *Proceedings of Machine Translation Summit X: Papers*. Phuket, Thailand, pp. 79–86. URL: https://aclanthology.org/2005.mtsummit-papers.11/ (cit. on p. 11).

Le Scao, Teven et al. (2023). "Bloom: A 176b-parameter open-access multilingual language model". In: (cit. on p. 13).

Lee, Minwoo et al. (2024). "Fine-grained Gender Control in Machine Translation with Large Language Models". In: *Proceedings of the 2024 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 5416–5430 (cit. on pp. 10, 21).

Levy, Shahar, Koren Lazar, and Gabriel Stanovsky (Nov. 2021). "Collecting a Large-Scale Gender Bias Dataset for Coreference Resolution and Machine Translation". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Ed. by Marie-Francine Moens et al. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 2470–2480. DOI: 10.18653/v1/2021.findings-emnlp.211. URL: https://aclanthology.org/2021.findings-emnlp.211/ (cit. on p. 11).

Lin, Xi Victoria et al. (2022). "Few-shot learning with multilingual generative language models". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9019–9052 (cit. on p. 13).

Liu, Danni and Jan Niehues (2023). "How Transferable are Attribute Controllers on Pretrained Multilingual Translation Models?" In: *arXiv preprint arXiv:2309.08565* (cit. on p. 12).

Lopez, Adam (2008). "Statistical machine translation". In: *ACM Computing Surveys (CSUR)* 40.3, pp. 1–49 (cit. on p. 6).

Lui, Marco and Timothy Baldwin (July 2012). "langid.py: An Off-the-shelf Language Identification Tool". In: *Proceedings of the ACL 2012 System Demonstrations*. Ed. by Min Zhang. Jeju Island, Korea: Association for Computational Linguistics, pp. 25–30. URL: https://aclanthology.org/P12-3005/ (cit. on p. 36).

Nădejde, Maria et al. (2022). "CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality". In: *arXiv preprint arXiv:2205.04022* (cit. on pp. 12, 13, 19, 20).

Ott, Myle et al. (June 2019). "fairseq: A Fast, Extensible Toolkit for Sequence Modeling". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Ed. by Waleed Ammar, Annie Louis, and Nasrin Mostafazadeh. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 48–53. DOI: 10.18653/v1/N19-4009. URL: https://aclanthology.org/N19-4009/ (cit. on p. 11).

Papineni, Kishore et al. (2002). "Bleu: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318 (cit. on p. 20).

Post, Matt (2018). "A call for clarity in reporting BLEU scores". In: *arXiv preprint arXiv:1804.08771* (cit. on p. 20).

Raffel, Colin et al. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140, pp. 1–67. URL: http://jmlr.org/papers/v21/20-074.html (cit. on p. 6).

Rarrick, Spencer et al. (2023). "Gate: A challenge set for gender-ambiguous translation examples". In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 845–854 (cit. on p. 10).

Raunak, Vikas et al. (2023). "Leveraging gpt-4 for automatic translation post-editing". In: *arXiv preprint arXiv:2305.14878* (cit. on pp. 11, 20).

Rei, Ricardo, José G. C. de Souza, et al. (Dec. 2022). "COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task". In: *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Ed. by Philipp Koehn et al. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 578–585. URL: https://aclanthology.org/2022.wmt-1.52/ (cit. on p. 21).

Rei, Ricardo, Craig Stewart, et al. (2020). "COMET: A neural framework for MT evaluation". In: *arXiv preprint arXiv:2009.09025* (cit. on pp. 20, 21).

Sánchez, Eduardo et al. (2023). "Gender-specific machine translation with large language models". In: *arXiv preprint arXiv:2309.03175* (cit. on p. 11).

Sarti, Gabriele et al. (2023). "RAMP: Retrieval and attribute-marking enhanced prompting for attribute-controlled translation". In: *arXiv preprint arXiv:2305.17131* (cit. on pp. 8, 12).

Saunders, Danielle, Rosie Sallis, and Bill Byrne (2020). "Neural Machine Translation Doesn't Translate Gender Coreference Right Unless You Make It". In: *arXiv preprint arXiv:2010.05332* (cit. on p. 10).

Schwenk, Holger et al. (Apr. 2021). "WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty. Online: Association for Computational Linguistics, pp. 1351–1361. DOI: 10.18653/v1/2021.eacl-main.115. URL: https://aclanthology.org/2021.eacl-main.115/ (cit. on p. 11).

Shazeer, Noam M. and Mitchell Stern (2018). "Adafactor: Adaptive Learning Rates with Sublinear Memory Cost". In: *ArXiv* abs/1804.04235. URL: https://api.semanticscholar.org/CorpusID:4786918 (cit. on p. 8).

Stanovsky, Gabriel, Noah A Smith, and Luke Zettlemoyer (2019). "Evaluating gender bias in machine translation". In: *arXiv preprint arXiv:1906.00591* (cit. on p. 10).

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le (2014). "Sequence to Sequence Learning with Neural Networks". In: *ArXiv* abs/1409.3215. URL: https://api.semanticscholar.org/CorpusID:7961699 (cit. on p. 6).

Tiedemann, Jörg (2012). "Parallel data, tools and interfaces in OPUS." In: *Lrec*. Vol. 2012. Citeseer, pp. 2214–2218 (cit. on p. 11).

Touvron, Hugo et al. (2023). "Llama 2: Open foundation and fine-tuned chat models". In: *arXiv preprint arXiv:2307.09288* (cit. on pp. 7, 22).

Üstün, Ahmet et al. (2024). "Aya model: An instruction finetuned open-access multilingual language model". In: *arXiv preprint arXiv:2402.07827* (cit. on p. 7).

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Neural Information Processing Systems*. URL: https://api.semanticscholar.org/CorpusID:13756489 (cit. on pp. 3, 6).

Werbos, Paul J. (1990). "Backpropagation Through Time: What It Does and How to Do It". In: *Proc. IEEE* 78, pp. 1550–1560. URL: https://api.semanticscholar.org/CorpusID:18470994 (cit. on p. 3).

Wu, Yonghui et al. (2016). "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation". In: *ArXiv* abs/1609.08144. URL: https://api.semanticscholar.org/CorpusID:3603249 (cit. on p. 6).

Xu, Haoran et al. (2023). "A paradigm shift in machine translation: Boosting translation performance of large language models". In: *arXiv preprint arXiv:2309.11674* (cit. on p. 7).

Xue, Linting et al. (June 2021). "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kristina Toutanova et al. Online: Association for Computational Linguistics, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41. URL: https://aclanthology.org/2021.naacl-main.41/ (cit. on p. 8).

Zhang, Biao et al. (2020). "Improving massively multilingual neural machine translation and zero-shot translation". In: *arXiv preprint arXiv:2004.11867* (cit. on p. 12).

A. Appendix

A.1. Prompts

For the target language Japanese, the LLM consistently generated romanized translations instead of utilizing the appropriate writing systems (Kanji, Hiragana, and Katakana) present in our dataset. To address this issue, we first translated our prompts into Japanese using Google Translate, ensuring they aligned with the expected linguistic characteristics. We then used these translated prompts to perform the translation tasks with the LLM.

The following is the prompt for formality controlled EN->JA translation.

The parameter attribute control in the first instruction will be set as formal (あなたは正式な形式を使用するプロの日本語翻訳者です。) or infomal (あなたは非公式な形式を使用するプロの日本語翻訳者です。).

`{Attribute control}` 翻訳にローマ字シ ステムを使用しないでください。 翻訳のみを提供してください。 追加の説明やメモは不要です。

正確に日本語に翻訳してください: `{sentence}`

We used the following prompts during the experiments. <attribute> is the placeholder for formality or gender control, i.e. "formal","informal","feminine", or "masculine". <target language> is the placeholder for the language we want to translate to. <Source Sentence> is the placeholder for the input sentence to be translated. <Current Translation" is placeholder for NLLB's translation result. <prompt examples> is placeholder for the translation examples from training dataset.

For Japanese, I set <attribute control> depending on the formality, where if the dataset is from "formal" form, <attribute control> stands for "あなたは正式な形式を使用するプロの日本語翻訳者です". If it is from "informal" form, <attribute control> stands for "あなたは非公式な形式を使用するプロの日本語翻訳者です。"

A.1.1. Llama query

A.1.1.1. Formality Zero-shot

You are a professional <target language>translator using the <attribute> form. ONLY provide the translation. NO additional explanations or notes.

Translate EXACTLY to <target language>: <Source Sentence>

A.1.1.2. Formality for Japanese Zero-shot

<attribute control>
翻訳にローマ字システムを使用しないでください。翻訳のみを提供してください。追加の説明やメモは不要です。

正確に日本語に翻訳してください : <Source Sentence>

A.1.1.3. Gender-Context Zero-shot

You are a professional <target language> translator. Only translate the text that appears AFTER the <sep> symbol. Use the text BEFORE <sep> solely to determine the correct gender form. Focus on the gender of a person. No additional explanations or notes.

Translate EXACTLY to <target language>: <Source Sentence>

A.1.1.4. Gender-Counterfactual Zero-shot

You are a professional <target language> translator focusing on <attribute> gender form. Maintain translation accuracy while ensuring <attribute> gender. No additional explanations or notes.

Translate EXACTLY to <target language>: <Source Sentence>

A.1.1.5. Formality Few-Shot

You are a professional <target language>translator using the <attribute> form.
Example Translations:
<prompt examples>
ONLY provide the translation. NO additional explanations or notes.

Translate EXACTLY to <target language>: <Source Sentence>

A.1.1.6. Formality for Japanese Few-Shot

<attribute control>
翻訳例 :
<prompt examples>
翻訳にローマ字システムを使用しないでください。翻訳のみを提供してください。追加の説明やメモは不要です。

正確に日本語に翻訳してください : <Source Sentence>

A.1.1.7. Gender-Context

You are a professional <target language> translator. Only translate the text that appears AFTER the <sep> symbol. Use the text BEFORE <sep> solely to determine the correct gender form. Focus on the gender of a person. No additional explanations or notes.
Example Translations:
<prompt examples>

Translate EXACTLY to <target language>: <Source Sentence>

A.1.1.8. Gender-Counterfactual Few-Shot

You are a professional <target language> translator focusing on <attribute> gender form. Maintain translation accuracy while ensuring <attribute> gender. No additional explanations or notes.

Example Translations:
<prompt examples>

Translate EXACTLY to <target language>: <Source Sentence>

A.1.2. Post-edit

A.1.2.1. Formality Zero-shot

Your task is to improve a given English-to-<target language> translation by ensuring that the translation correctly reflects the <attribute> references and grammar. Focus specifically on correcting pronouns, possessive forms, and any <attribute>-related grammatical structures. Provide translation only.

Source Sentence: <Source Sentence>
Current Translation:<Current Translation>

A.1.2.2. Gender-Counterfactual Zero-shot

Your task is to improve the given <target language> translation. Review and correct the gender-related elements, ensuring the main subject/actor reflects <attribute> gender. Also improve the overall translation quality for accuracy. Convert pronouns, possessives, and grammatical agreements related to the main subject to match the <attribute> form.
Source Sentence: <Source Sentence>
Current Translation:<Current Translation>

A.1.2.3. Gender-Context Zero-shot

You are a professional English-to-<target language> translator. Your task is to improve the given translation pair. First, analyze the context (the part before <sep> in the source sentence) to determine the gender. Then use this gender information to improve the given translation of the main part (the part after <sep> in the source sentence). Ensure the improved translation maintains high-quality fluency as the given translation. Output only the improved translation of the part after <sep>.

Source Sentence: <Source Sentence>
Current Translation:<Current Translation>

A.1.2.4. Formality Few-shot

Your task is to improve a given English-to-<target language> translation by ensuring that the translation correctly reflects the <attribute> references and grammar. Focus specifically on correcting pronouns, possessive forms, and any <attribute>-related grammatical structures. Provide translation only.
Example Translations:
<prompt examples>

Source Sentence: <Source Sentence>
Current Translation:<Current Translation>

A.1.2.5. Gender-Counterfactual Few-shot

our task is to improve the given <target language> translation. Review and correct the gender-related elements, ensuring the main subject/actor reflects <attribute> gender. Also improve the overall translation quality for accuracy. Convert pronouns, possessives, and grammatical agreements related to the main subject to match the <attribute> form. Example Translations:
<prompt examples>

Source Sentence: <Source Sentence>
Current Translation:<Current Translation>

A.1.2.6. Gender-Context Few-shot

You are a professional English-to-<target language> translator. Your task is to improve the given translation pair. First, analyze the context (the part before <sep> in the source sentence) to determine the gender. Then use this gender information to improve the given translation of the main part (the part after <sep> in the source sentence). Ensure the improved translation maintains high-quality fluency as the given translation. Output only the improved translation of the part after <sep>.
Example Translations:
<prompt examples>

Source Sentence: <Source Sentence>
Current Translation:<Current Translation>

A.1.3. LLM as Gender Evaluator

A.1.3.1. 0-10 Scale rating

You are evaluating an English-to-<target language> translation for <attribute> gender accuracy. Focus ONLY on how accurately the translation reflects the gender of the PERSON mentioned in the sentence. Rate the accuracy on a scale of 0 to 10, where 0 is completely inaccurate and 10 is perfectly accurate. Provide the result in this exact format: 'gender accuracy score: [0-10].' ONLY provide the score. NO ADDITIONAL EXPLANATION.

Rate the following translation for <attribute> gender accuracy:
Source Sentence: <Source Sentence>
Current Translation: <Current Translation>

A.1.3.2. Binary rating

This is an English-to-<target language> translation classification task. Your goal is to evaluate if the translation uses the correct gender and gender-related grammar. If the translation is

accurate regarding gender, return 'Accurate'. Otherwise, return 'Inaccurate'.

Classify the following translation for gender control:
Source Sentence: <Source Sentence>
Current Translation: <Current Translation>

A.1.4. Multi-language prompt

You are a professional <target language> translator. The following are translations from English to <example language> in <attribute> tone.

Example Translations:
<prompt examples>

Translate to <target language> in <attribute> tone: <sentence>

A.1.5. Identify Missing Information

A.1.5.1. Binary rating

You are evaluating an English-to-<target language> translation. Your task is to determine if additional information is required to translate the sentence. Respond with 'yes' if additional information is necessary, or 'no' if it is not. No explanation or comments.

Source Sentence: <Source Sentence>
Does the translation require additional context?

A.1.5.2. 0-10 Scale rating

You are evaluating an English-to-<target language> translation. Your task is to assess how much ADDITIONAL gender information the source sentence needs for accurate translation.
Important: Only evaluate gender clarity for the MAIN SUBJECT/ACTOR of the sentence. If the sentence mentions other people (like references, examples, or comparisons), ignore their gender clarity".
Check for explicit gender indicators, such as: Pronouns ('he', 'she', 'his', 'her') , Gendered terms ('man', 'woman', 'father', 'mother') , Possessive adjectives ('his book', 'her idea')."
Rate on a scale of 0 to 10. If the source sentence already provides sufficient clarity and no additional gender information is needed, rate it as 0. If the source sentence is highly ambiguous and additional gender details are essential for correct translation, then rate it as 10. ONLY provide the rating score. NO ADDITIONAL EXPLANATION.

Source Sentence: <source sentence>
Rate, how necessary is additional gender information for accurate translation?