



A Study on Ethical Issues and Mitigation Strategies in Translation with Large Language Models

Chang Wei¹

¹*School of Foreign Languages and Cultures, Panzhihua University, Sichuan province, China*
Corresponding Author: Chang Wei

Date of Submission: 02-02-2025

Date of Acceptance: 12-02-2025

Abstract: With the widespread application of large language models (LLMs) in the field of translation, ethical issues arising from their use have gradually attracted attention. This paper aims to conduct an in-depth analysis of the ethical issues in translation with LLMs and propose effective mitigation strategies. By examining the technical principles of LLMs and translation ethics theory, this study provides a detailed analysis of the ethical issues and their causes in LLM translation, focusing on accuracy and faithfulness, cultural and value dissemination, user privacy, and data security. A range of mitigation strategies are proposed from three aspects: data, model, and post-processing, including data collection and cleaning, model improvement and optimization, as well as human review and multi-model integration. This research not only enriches the theoretical studies on the application of LLMs in translation but also offers practical guidance for the optimization of real translation systems. It also suggests that future research could be directed towards interdisciplinary integration and exploration of new application scenarios.

Keywords: Large Language Model Translation, Translation Ethics, Mitigation, Strategies

I. Introduction

1.1 Research Background

1.1.1 Widespread Application and Impact of Large Language Models in Translation

Large language models (LLMs), as cutting-edge technologies in the field of natural language processing (NLP), have demonstrated extensive application prospects and profound influence in translation. From a theoretical perspective, LLMs are based on the Transformer architecture and employ unsupervised learning on vast amounts of text data to construct a large and complex system of linguistic knowledge. This endows them with powerful capabilities in language understanding and generation, providing robust technical support for translation tasks.[1]

In terms of application scenarios, LLMs have

been deeply integrated into various fields. In the translation of scientific and technological literature, they can rapidly process specialized terminology and complex sentence structures, helping researchers overcome language barriers to access cutting-edge international research findings in a timely manner. This, in turn, promotes global scientific cooperation and knowledge sharing. In the realm of business translation, LLMs can efficiently translate business documents such as contracts and reports, meeting the needs of enterprises' global operations and enhancing the efficiency and accuracy of business activities. Additionally, in everyday communication and travel translation scenarios, real-time translation tools leveraging LLMs enable instant language conversion, significantly facilitating communication among speakers of different languages.

Regarding their impact on the translation industry, LLMs have driven transformation and innovation. [2] On one hand, they have lowered the barriers to translation, allowing non-professional translators to complete basic translation tasks with the aid of these tools, thereby altering the traditional translation market landscape. On the other hand, they have prompted professional translators to enhance their capabilities, shifting from mere language converters to language experts and cultural consultants. This shift places greater emphasis on in-depth quality control of translations and the accurate conveyance of cultural connotations.

From an academic research standpoint, LLMs have provided new perspectives and methodologies for translation studies. [3] Scholars can utilize translation data generated by LLMs to conduct in-depth research on language conversion patterns, cultural adaptability in translation, and translation quality assessment indicators, thereby advancing the development of translation theory and practice. Meanwhile, the application of LLMs in translation has also sparked a series of academic discussions, such as translation ethics, data privacy protection, and model interpretability. These discussions have created new opportunities for interdisciplinary research.

1.1.2 Potential Harms of Ethical Issues in



Translation

Despite the remarkable achievements of LLMs in translation, the potential harms of their ethical issues cannot be overlooked. In terms of accuracy and faithfulness, models trained based on statistical patterns and large-scale data often struggle with texts that are semantically and culturally complex. For example, when translating specific cultural images in ancient literary works, errors are likely to occur, which can undermine the artistic value of the works and hinder cross-cultural communication. Regarding the dissemination of culture and values, cultural biases in the training data can lead to translations that reinforce the dominance of certain cultures while distorting others. This is evident in the translation of traditional festivals of different ethnic groups, where an imbalance in the representation of Eastern and Western festivals can cause cultural misunderstandings and conflicts, thereby disrupting the global cultural ecology. In terms of user privacy and data security, the training of models involves a large amount of sensitive data that may contain privacy information. In the absence of robust security and privacy protection mechanisms, data breaches can infringe upon individual rights, lead to legal disputes, and erode user trust. [4]

1.2 Research Objectives and Significance

1.2.1 Research Objectives

The objective of this study is to thoroughly investigate the various ethical issues arising from the translation process of large language models. From the ethical dimensions of accuracy and faithfulness, cultural and value dissemination, and user privacy and data security, this study meticulously combs through the roots, manifestations, and mechanisms of these issues. Additionally, from multiple perspectives including technical optimization, data management, and ethical review, this study proposes feasible ethical guidelines. Technically, this study explores the improvement of model architecture and algorithms to enable better understanding and processing of texts that are semantically complex and rich in cultural connotations. In terms of data management, scientific methods for data collection, cleaning, and annotation are established to ensure the comprehensiveness, accuracy, and unbiasedness of training data. At the level of ethical review, a rigorous review mechanism is established to conduct ethical assessments of model translation results, identifying and correcting potential ethical issues in a timely manner.

1.2.2 Research Significance

Theoretically, this study enriches the theoretical research on large language models in the

field of translation. Currently, research on LLM translation primarily focuses on technical applications and performance enhancements, with relatively weak emphasis on ethical issues. This study fills this theoretical gap by providing a systematic ethical analysis framework and methods for bias research, promoting the interdisciplinary integration of natural language processing with ethics, sociology, and other disciplines. It expands the theoretical boundaries of LLM research and injects new vitality into the development of translation theory.

From a practical standpoint, this study holds significant value for the translation industry and social development. In the translation industry, it helps translation practitioners and related enterprises better understand and address the issues in LLM translation, thereby improving translation quality and service levels and enhancing market competitiveness. For society, it promotes fair, objective, and harmonious cross-cultural communication, reducing cultural misunderstandings and social conflicts caused by ethical issues in translation. It also helps maintain social fairness and harmony and promotes the equal exchange and common development of diverse cultures in the process of globalization. For example, in international business communication, cultural dissemination, and academic cooperation, ensuring accurate translation content provides strong support for the smooth conduct of various activities.

II. Related Theoretical Foundations and Research Status

2.1 Overview of Large Language Models

2.1.1 Technical Principles and Architecture

The Transformer architecture is the core foundation of large language model translation. [5] It abandons the sequential processing methods of traditional Recurrent Neural Networks (RNNs) and Long Short - Term Memory networks (LSTMs), and adopts the Multi - Head Attention mechanism, which can process the information of each position in the input sequence in parallel. This enables the model to more efficiently capture global dependencies when dealing with long texts, greatly improving computational efficiency and language comprehension ability. For example, when translating long sentences, traditional models may make mistakes due to difficulty in remembering previous information, while the Transformer architecture can pay attention to all parts of the sentence simultaneously and accurately grasp semantic associations.

The attention mechanism is a key innovation of the Transformer architecture. [6] Its core idea is to enable the model to dynamically allocate the degree of attention to different parts of the input text during the



translation process. In the calculation process, the model calculates the correlation scores between each position in the input sequence and the target position according to the content to be translated, and then performs a weighted sum of the input information based on these scores. For instance, when translating the English sentence "I bought a book which was written by a famous author yesterday." into Chinese, when the model processes the relative clause "which was written by a famous author", it will focus on the information related to "book", thus accurately translating it as "我昨天买了一本由著名作家写的书" to ensure the accurate conveyance of the modifying relationship.

Large language models typically adopt the training paradigm of Pre-training and Fine-tuning. In the pre-training stage, the model is trained based on large-scale unsupervised text data. It learns the general patterns and semantic knowledge of language through tasks such as predicting the next word, and builds strong language understanding and generation capabilities. For example, GPT-3 used a vast amount of Internet text during pre-training. In the fine-tuning stage, for the translation task, the model is further trained on a specific translation dataset, and the model parameters are adjusted to meet the specific requirements of translation, thereby improving the performance in the translation task.

In the translation task, large language models often adopt the Encoder-Decoder structure. The encoder is responsible for transforming the source-language text into an intermediate semantic representation, and this process extracts the key features and semantic information of the text. The decoder, based on the output of the encoder, combines the grammatical and semantic rules of the target language to gradually generate the target-language text. Taking neural machine translation as an example, the encoder encodes the source-language sentence into a fixed-length vector, and the decoder starts from this vector and generates the words of the target language one by one, ultimately completing the translation of the entire sentence and achieving the transformation from the source language to the target language.

2.1.2 Training Data and Training Methods

Training data serves as the cornerstone of the translation capabilities of large language models, and it has a wide range of sources. On one hand, a large number of publicly available text corpora are important data sources, such as Wikipedia, news reports, academic papers, etc. These data cover rich domain knowledge, language expressions, and

cultural background information, enabling the model to learn general language patterns and semantic understanding. For example, the multilingual versions of Wikipedia contain professional knowledge in different fields. By learning these contents, the model can improve its ability in translating professional terms, etc. On the other hand, domain-specific datasets are crucial for enhancing the model's translation performance in specialized fields. For instance, medical literature and case data in the medical field, and legal regulations and case data in the legal field. These data enable the model to learn the language characteristics and professional vocabulary of specific fields, enhancing the accuracy and adaptability in professional translation tasks.

The diversity of training data is also of great significance. It includes not only data from different fields but also texts in multiple language pairs. Abundant language-pair data allows the model to learn the differences in grammatical structures, vocabulary correspondences, and the conversion methods of cultural connotations between different languages. For example, data containing multiple language pairs such as Chinese-English, Chinese-French, and Chinese-Russian can enable the model to perform better in multilingual translation tasks and avoid translation limitations caused by single-type data.

High-quality training data is one of the key factors ensuring the model's performance. Data quality issues include the accuracy, consistency, and integrity of the data. [7] Inaccurate data, such as texts with spelling errors, grammatical mistakes, or incorrect annotations, can mislead the model's learning and lead to translation errors. Inconsistent data, for example, when the same concept has different expressions in different texts without unified processing, can cause the model to learn ambiguous or incorrect language patterns. Incomplete data may prevent the model from comprehensively learning language knowledge, affecting its generalization ability.

To improve data quality, strict data preprocessing is required. [8] Preprocessing steps typically include data cleaning, deduplication, tokenization, and annotation, etc. Data cleaning aims to remove noisy data, such as deleting garbled characters and invalid characters. The deduplication operation avoids the redundant impact of duplicate data on model training. Tokenization divides the text into the smallest semantic units for easy processing by the model. Annotation adds additional information to the data, such as part-of-speech tagging and syntactic annotation, helping the model better understand the text structure and semantics. For example, when



training a Chinese - English translation model, tokenizing Chinese texts and performing part of speech tagging on English texts can improve the model's understanding and translation capabilities of the two languages.

In the training process of large language model translation, optimization algorithms play a crucial role. Common optimization algorithms such as Stochastic Gradient Descent (SGD) and its variants Adagrad, Adadelata, Adam, etc., are widely used. SGD updates the model parameters by calculating the gradient of each training sample. However, due to its fixed update step size, it is prone to getting trapped in local optimal solutions in complex loss - function spaces. Adagrad adaptively adjusts the learning rate according to the gradient history of each parameter, which can effectively handle sparse data, but may suffer from the problem of premature decay of the learning rate. Adadelata improves on Adagrad by using the cumulative sum of the squared gradients to dynamically adjust the learning rate, avoiding excessive decay of the learning rate. The Adam algorithm combines the advantages of Adagrad and Adadelata, taking into account both the first - moment and second - moment estimates of the gradient. In practical applications, it exhibits good performance, can converge to a better solution more quickly, and improve the training efficiency and translation quality of the model.

Hyperparameter tuning is an important part of large language model training. Hyperparameters are parameters that need to be manually set before the training starts, such as the learning rate, batch size, number of hidden - layer neurons, etc. The choice of these parameters has a significant impact on the model's performance. The learning rate determines the step size of parameter updates during the training process of the model. A too - large learning rate may cause the model to fail to converge during training or even diverge. A too - small learning rate will make the training process extremely slow, increasing the training time and computational resource consumption. The batch size refers to the number of samples input to the model in each training. An appropriate batch size can balance the stability of training and computational efficiency. The number of hidden - layer neurons affects the model's expressive ability. Too many neurons may lead to overfitting of the model, while too few may prevent the model from learning sufficient features.

Hyperparameter tuning usually employs methods such as grid search, random search, and Bayesian optimization. Grid search exhaustively searches all possible combinations in the specified hyperparameter space to find the optimal solution, but

it has a high computational cost. Random search randomly samples in the hyperparameter space for experimentation, which can reduce the computational cost to a certain extent, but may not find the global optimal solution. Bayesian optimization uses Bayes' theorem to estimate the posterior distribution of hyperparameters, and constructs a surrogate model to guide the selection of hyperparameters, enabling a more efficient search of the hyperparameter space, and gradually gaining favor in practical applications.

2.2 Translation Ethics Theory

2.2.1 Definition and Scope of Translation Ethics

Translation ethics is a relatively new concept in translation studies, first proposed by French scholar Antoine Berman in 1984. It refers to the relationships between the subjects, objects, environment, society, and culture involved in translation, as well as the behavioral norms for handling these relationships.[9] Translation ethics is closely related to the morality of translators, but it focuses more on the understanding of the relationships between the subjects and objects of translation activities and the socio-cultural system, while translator morality emphasizes the consensus value standards for translators to handle various relationships involved in translation. Translation ethics serves as the foundation for translator morality, which in turn maintains and adjusts the ethical relationships involved in translation. Specifically, translation ethics is concerned with how translators balance fidelity to the original text and the needs of the target audience, how they handle the impact of cultural differences and power relations on translation, and how they maintain the professionalism and integrity of translation. It is not merely a constraint on translation behavior but also a value orientation guiding translation activities towards positive and beneficial directions.

In a broader sense, translation ethics involves five types of relationships. First, the relationship between the translator and the original author. The ethical relationship between the translator and the original author is the cornerstone of translation ethics. The principle of fidelity runs through the entire process, requiring the translator to accurately reproduce the semantics and style of the original text from the lexical to the discourse level. The translator must not only match the language form but also deeply understand the author's intentions, respect their unique expression within the cultural context, and avoid misinterpretation and misunderstanding to ensure that the original ideas and emotions are conveyed intact. Second, the relationship between the translator and the target audience. Facing the target audience, the translator bears the important responsibility of



building an effective bridge for communication. On one hand, the translation should be readable, matching the audience's language habits and knowledge level. On the other hand, the translator must not sacrifice the accuracy of information for the sake of simplicity, ensuring that the audience receives the same amount of true content as the original text to achieve smooth and error-free cross-linguistic communication. Third, the relationship between the translator and the translation client. In cooperation with the translation client, the spirit of contract is crucial. The translator strictly adheres to the contract, delivering the translation results on time and with high quality, without delay or negligence. At the same time, the translator keeps confidential any confidential information learned during the translation process, maintaining the client's commercial interests and trust with integrity and responsibility. Fourth, the relationship of cultural dissemination and exchange. In terms of cultural dissemination, translation ethics emphasizes cultural equality and respect. The translator abandons cultural prejudices, treats different cultures with an inclusive heart, and promotes the exchange and collision of diverse cultures. Actively spreading valuable cultural content, the translator becomes a promoter of cultural mutual learning, contributing to the prosperity, development, and coexistence of cultures. Fifth, the relationship between professional ethics and industry standards. From the perspective of professional ethics, the translator continuously improves professional skills, keeps pace with industry development, enriches knowledge reserves through learning, and refines translation skills. Strictly adhering to industry standards, whether in quality control, term usage, or copyright protection, the translator follows the rules to shape a good image of the translation industry and maintain its healthy ecosystem.

2.2.2 Differences between Traditional and Modern Translation Ethics

Traditional translation ethics primarily revolves around the principle of fidelity, focusing on the relationship between the translator and the original author.[10] For example, Yan Fu, in his translation of Western classics, exemplifies this approach. When translating *Evolution and Ethics*, he maintained a high degree of loyalty to the original text, striving to accurately convey Huxley's ideas. In terms of language, Yan Fu endeavored to preserve the structure and terminology of the original text. The core concept of "survival of the fittest through natural selection" was translated in a concise manner that remained faithful to the original, allowing Chinese readers to directly engage with advanced Western ideas while

retaining the cultural imagery of the source text. This demonstrated the principle of "faithfulness" and reflected an inviolable respect for the original work.

When addressing the target audience, traditional translation ethics, while also pursuing the transmission of information, emphasized that the translation should conform to the grammatical rules and basic expression habits of the target language. However, it paid less attention to aligning with the audience's cultural background and reading habits. For instance, in the early translation of Buddhist scriptures, translators focused mainly on accurately converting Sanskrit texts into Chinese, adhering to the basic grammatical rules of Chinese, but paid little attention to the comprehension level and cultural background differences of the general public, resulting in some translations being abstruse and difficult to understand. Regarding translation clients, traditional concepts relied more on verbal agreements or simple contracts, requiring translators to deliver on time and ensure basic translation quality. For example, in ancient times, the translation of poetry and prose among literati was often based on verbal commitments, with relatively lenient requirements for delivery time and quality.

In terms of cultural dissemination, traditional translation ethics primarily aimed to spread the culture of the source language, with less consideration for equal interaction between cultures. For example, when Matteo Ricci translated Western scientific works, his main goal was to introduce Western astronomical and mathematical knowledge into China, focusing on the dissemination of Western culture with little consideration of the reciprocal influence of Chinese culture on Western culture. In terms of professional ethics, the requirements for translators' professional abilities were centered on language conversion skills, with relatively loose industry standards. For instance, ancient diplomatic translators mainly relied on their linguistic talent and experience, lacking unified regulatory constraints.

With the development of the times, modern translation ethics has become more diverse and complex. In the relationship between the translator and the original author, the focus is no longer on absolute fidelity but on flexible handling based on an understanding of the author's intentions, aiming for functional equivalence.[11] For example, in the translation of contemporary literary works, when translating Mo Yan's novels, translators will adapt certain culturally specific expressions according to the reading habits of Western audiences. While preserving the core ideas and artistic style of the work, they make it more accessible to Western readers.

When addressing the target audience, modern translation ethics places a high priority on meeting the



needs of the audience. For example, in the translation of children's literature, translators will fully consider the cognitive level and reading preferences of children, using simple, lively, and easily understandable language to transform classic stories into versions suitable for children, ensuring that the information is effectively received. When collaborating with translation clients, modern translation ethics emphasizes the rigor and legal binding nature of contracts, clarifying the rights and obligations of both parties, and imposing stricter requirements on the translator's confidentiality responsibilities. In commercial translation projects, translators and clients sign detailed contracts specifying the translation content, delivery time, quality standards, and confidentiality clauses. Breach of contract entails legal liability.

In the realm of cultural dissemination and exchange, modern translation ethics advocates cultural equality and promotes two-way interaction and integration between different cultures. Translators are encouraged to explore cultural commonalities and eliminate misunderstandings. For example, in the translation of subtitles for film and television works, when translating foreign movies, translators skillfully handle cultural references, making them understandable to domestic audiences while also incorporating elements of their own culture to facilitate cultural exchange. In terms of professional ethics and industry standards, modern translation ethics requires translators to engage in continuous learning and acquire interdisciplinary knowledge. Industry standards have also become more refined, covering various aspects such as translation processes, quality assessment, and terminology management, to ensure the healthy development of the translation industry. Today, in the field of medical translation, translators are required not only to master medical terminology but also to be aware of the latest medical research findings. The industry ensures translation quality through the development of unified terminology lists and quality assessment standards.

III. Ethical Issues in Translation with Large Language Models

3.1 Ethical Issues of Accuracy and Fidelity

3.1.1 Ethical Risks Caused by Information Loss and Incorrect Translation

Information loss and incorrect translation by large language models (LLMs) can trigger a series of ethical issues, which have already been reflected in practical applications. First, LLMs may inadvertently disclose private or sensitive information during information processing and generation, leading to

privacy and data security issues. For example, in April 2023, ChatGPT was reported to have a severe leakage issue, where some users could access other users' names, email addresses, chat record titles, and the last four digits of credit card numbers. This kind of information leakage not only violates user privacy but may also lead to identity theft and other security problems. [12] Second, the training of LLMs relies on large datasets, which often originate from the real world and may carry existing social inequalities and biases. This can result in the models reflecting unjust social concepts when processing information, potentially marginalizing vulnerable groups or inciting hatred and violence.[13] For instance, biases in the datasets, such as discrimination based on race or socioeconomic status, may be inadvertently reinforced and solidified through the "technological neutrality" of LLMs. In translation, this can lead to improper descriptions or unfair treatment of certain groups, causing ethical controversies.

Moreover, the risks associated with information processing and output by LLMs, and the resulting harm, are urgent issues that need to be addressed in the field of artificial intelligence. For example, the content learned by models during training may be limited and biased by the training data, leading to discrepancies between generated content and facts. In translation, this can result in the spread of incorrect information, misleading the public's understanding and judgment of events. In professional fields such as medicine and law, where the accuracy and completeness of information are crucial, information loss and incorrect translation by LLMs can have severe consequences. For example, incorrect translation in medical translation may lead to misdiagnosis, mistreatment, and even endanger patients' lives. In the legal field, the translation of legal texts requires high precision, and any information loss or error may lead to legal disputes, affecting the fairness and authority of the law. Additionally, if LLMs make errors or lose important information when translating academic literature, it may lead to inaccurate and incomplete academic research. Researchers may base their studies on incorrect translations, leading to erroneous conclusions, which not only affect the quality of academic research but may also lead to academic misconduct, such as plagiarism. Incorrect translation may also hinder the dissemination and exchange of academic ideas, affecting the overall development of the academic community.

The translation results of LLMs may also impact social values. If incorrect translation or information loss leads to the spread of incorrect concepts or values, it may challenge mainstream



social values. [14] For example, models may spread content with biases or discrimination during translation, misleading the public's value orientation and affecting the moral climate of society. The existence of these issues reminds us to be cautious when using LLMs and to take corresponding measures to reduce these ethical risks.

3.1.2 Distortion of Source Language Culture and Author's Intent

Distortion of source language culture and author's intent during translation by LLMs can also trigger a series of ethical issues. First, cultural distortion mainly manifests as misunderstandings or neglect of the cultural background of the source language. Humor, irony, or teasing expressions in certain cultures may be misunderstood as disrespect or offense in other cultures, resulting in translation outcomes that do not align with the cultural connotations of the original text. [15] This kind of cultural misreading not only affects the effectiveness of cross-cultural communication but may also lead to cultural conflicts. Moreover, due to the limitations of training data, LLMs may fail to accurately understand specific customs or values in the source language, leading to the loss or misunderstanding of cultural information during translation. For example, in the initial version of an AI translation tool developed by a tech company, insufficient understanding of certain culturally symbolic terms led to incorrect translations, causing misunderstandings of the source language culture among target language readers. This cultural distortion not only affects the effectiveness of cross-cultural communication but may also have a negative impact on the dissemination of the source language culture.

Second, distortion of the author's intent can also raise ethical issues. LLMs may misunderstand or inaccurately handle the original text during translation, resulting in translations that fail to accurately reflect the author's emotions, thoughts, and cultural background. For example, in literary translation, it is essential for translators to accurately convey the author's intent; otherwise, readers' understanding of the work may be biased. In some cases, LLMs may oversimplify or misunderstand the original text, omitting key information or providing incorrect interpretations. For instance, when translating Pearl S. Buck's English version of *Water Margin*, the translator's insufficient understanding of the cultural background of the original text may lead to translations that fail to accurately convey the author's intent. This distortion not only affects readers' understanding of the work but may also have a negative impact on the author's reputation and the

value of the work. [16] Therefore, when using LLMs for translation, it is essential to pay special attention to the accurate conveyance of cultural context and the author's intent to reduce ethical risks.

3.2 Ethical Issues in the Dissemination of Culture and Values

3.2.1 Cultural Hegemony and the Spread of Cultural Bias

Large language models (LLMs) may lead to cultural dominance during the translation process due to linguistic advantages. These models are primarily trained on corpora based on mainstream languages such as English, which results in an overemphasis on English expressions and grammatical rules in multilingual text translation. This, in turn, affects the natural expression of other languages and the accurate conveyance of their cultural connotations. Additionally, the training data of LLMs often contains values and concepts from specific cultural contexts, which may be disseminated indiscriminately during translation while ignoring the uniqueness and diversity of other cultures. For example, the individualistic values of Western culture may be overemphasized, while the values of collectivist cultures are underrepresented. Moreover, since the training data mainly originates from mainstream cultures, there is a relative scarcity of materials from minority cultures, local cultures, and subcultures. This may lead to the neglect, misunderstanding, or even marginalization of these minority cultures during translation. For instance, when translating texts involving minority languages and cultures, models may fail to accurately understand and convey their unique cultural connotations.

The training data of LLMs may contain stereotypes of certain cultural groups, which may be unconsciously reinforced during translation, leading to misunderstandings and biases towards specific cultures. For example, when translating texts about a particular country or region, some stereotypical words or expressions may be used. Due to the lack of in-depth understanding and sensitivity to different cultural backgrounds, LLMs may produce cultural misunderstandings during translation, resulting in inaccurate translations. Specific customs, beliefs, or values of certain cultures may be distorted or omitted during translation. Furthermore, LLMs may tend to apply a universal cultural standard when translating texts, thereby ignoring the differences and diversity between cultures. This may lead to translation results that lack cultural adaptability and fail to meet the needs of different cultural audiences.

3.2.2 Inappropriate Influence on the Target



Language Culture

LLMs may have inappropriate impacts on the target language culture in various ways during the translation process. First, the training data of these models often comes from the real world, which may contain existing social inequalities and biases. These biases may be inadvertently disseminated during translation, leading to misunderstandings or distortions of the target language culture. [17] For example, the model may incorporate stereotypes of certain cultural groups into the translation content, thereby reinforcing these inaccurate notions. Such dissemination may not only affect the image of the target language culture but also cause emotional harm to specific groups. Second, when translating, LLMs may output based on the cultural values reflected in their training data, which may differ from the values of the target language culture. For instance, views on gender, family, or social relationships in certain cultures may differ from the concepts learned by the model, resulting in translation content that does not align with the actual values of the target culture. This conflict may lead to misunderstandings of the translation content by the audience of the target language culture and may even trigger cultural conflicts.

When handling translation tasks, LLMs may tend to adopt a universal cultural standard, thereby ignoring the uniqueness and diversity of the target language culture. [18] This tendency may lead to translation results that lack cultural adaptability and fail to meet the needs of different cultural audiences. For example, the model may fail to accurately convey specific customs, beliefs, or values of the target language culture and may even omit these important elements. Moreover, the translation output of LLMs may have a certain degree of homogeneity. This trend towards homogenization may have a "flattening" effect on the target language culture. The model may simplify or omit some culturally distinctive expressions during translation, thereby weakening the richness and uniqueness of the target language culture. This homogenization may not only affect the inheritance and development of the target language culture but also have a negative impact on cross-cultural communication.

Due to the lack of in-depth understanding and sensitivity to different cultural backgrounds, LLMs may produce cultural misunderstandings during translation, resulting in inaccurate translations. For example, specific customs, beliefs, or values of certain cultures may be distorted or omitted during translation. Such inaccurate translations may mislead the audience of the target language culture and cause them to have incorrect perceptions of the source

language culture. Moreover, the inappropriate influence of LLMs on the target language culture during translation raises ethical issues, especially in terms of responsibility attribution. Since the output of the model is determined by its training data and algorithms, it is difficult to clearly define who should be held responsible for these inappropriate impacts. This may lead to a lack of an effective accountability mechanism when dealing with issues such as cultural bias and misunderstanding.

IV. Strategies for Mitigating Ethical Issues in Translation with Large Language Models

4.1 Data Processing

4.1.1 Diversification of Data Collection

Data collection is the cornerstone of training large language models (LLMs), and its diversity directly determines whether the model can learn comprehensive and objective language patterns.[19] In practice, it is essential to widely cover texts from different cultural, racial, gender, and age groups. For example, in the field of news, attention should not be limited to a few mainstream media outlets but should include news content from media with diverse political stances and cultural backgrounds around the world. News reports from the Middle East, for instance, can expose the model to unique cultural contexts and language habits, avoiding cultural misunderstandings when translating texts related to this region. In terms of literary works, in addition to classic masterpieces, innovative expressions and unique perspectives from emerging literary schools can enrich the model's language material repository. When collecting academic papers, it is necessary to span multiple disciplines, including natural sciences, social sciences, and humanities. The professional terminology and logical expressions from different disciplines can enhance the model's ability to handle complex texts. When collecting data from social media, attention should be paid to the discourse of different regional and interest groups. For example, the internet slang from gaming enthusiast communities and unique vocabulary in food lovers' shares can enable the model to learn more life-oriented and diverse language expressions, reducing biases caused by homogeneous data.

4.1.2 Data Cleaning and Annotation Review

Collected data, like unpolished jade, must undergo rigorous cleaning and annotation review to become high-quality material for model training. When cleaning data, a combination of advanced text analysis technologies and manual screening should be employed to remove data containing biases,



discriminatory language, or incorrect information.[20] For example, when processing social media data, natural language processing tools can be used to identify and filter out aggressive and discriminatory comments, followed by a secondary manual review to ensure the purity of the data. In the data annotation phase, the competence of professional annotators is crucial. For the translation annotation of vocabulary involving different cultural customs, annotators must not only be proficient in both languages but also have an in-depth understanding of the relevant cultural backgrounds. For example, when annotating translations of traditional Chinese festival-related terms, annotators need to be clear about the historical origins and celebration methods of each festival to accurately annotate their connotations and avoid incorrect translations due to cultural differences. If annotations with ethical issues are detected, an immediate correction or deletion process should be initiated to provide high-quality, ethically sound data for model training.

4.1.3 Strengthening Privacy Protection

In today's highly interconnected information environment, protecting user data privacy is the baseline for the application of large language models. Adopting advanced privacy protection technologies is key to achieving this goal.[21] Differential privacy technology adds appropriate noise to data to cleverly conceal individual sensitive information. Even if an attacker gains access to part of the data, it is difficult to extract valuable personal privacy information. Homomorphic encryption technology allows computations to be performed on ciphertext, ensuring that data remains encrypted throughout the processing, thus safeguarding data confidentiality. At the same time, during data usage, it is necessary to clearly define the scope and purpose of data use and strictly comply with relevant laws and regulations, such as the European Union's General Data Protection Regulation (GDPR). Data can only be used with the explicit consent of users, and users must be clearly informed about the purpose and protection measures of data usage to effectively safeguard user privacy and security.

4.2 Model Design and Training

4.2.1 Incorporating Ethical Constraints

Integrating a dedicated ethical constraint module into the model architecture design is an important measure to avoid ethical issues at the source.[22] For example, using a penalty term in the loss function, when the model output contains biased or discriminatory content, the model is punished by increasing the loss value, prompting it to adjust its

learning direction. For instance, when translating texts involving gender-specific job descriptions, if the model outputs translations with gender stereotypes, such as translating "male nurse" into a derogatory expression, the loss function penalty term will take effect, increasing the model's loss value. In subsequent training, the model will attempt to adjust its parameters to gradually master a fair and unbiased language translation pattern. This constraint mechanism can supervise and guide model behavior in real-time during training, effectively preventing the emergence of ethical issues.

4.2.2 Iterative Training and Optimization

Iterative training and optimization are essential for improving the ethical performance of models. After each round of training, a comprehensive ethical assessment of the model's output should be conducted using a variety of evaluation metrics. In addition to traditional translation quality assessment metrics such as BLEU (Bilingual Evaluation Understudy), it is also necessary to introduce metrics specifically targeting ethical issues.[23] Adversarial training is also an effective method to enhance the model's resistance to ethical issues. By carefully designing adversarial samples and having the model engage in adversarial interactions with them, for example, constructing adversarial samples containing various bias scenarios, the model can enhance its ability to identify and respond to various ethical risks as it continuously confronts these challenges.

4.2.3 Human Intervention and Guidance

Timely human intervention during training can inject correct values and language understanding into the model. For words or phrases with culturally specific meanings that are prone to ambiguity, accurate translation examples should be provided manually, along with detailed explanations of their cultural background and correct usage. For example, the term "kung fu" is widely accepted in English as "Kung Fu," but without explaining its rich cultural connotations related to Chinese martial arts, the model may fail to convey its essence accurately when translating relevant texts. Through human intervention, the model can better understand and handle these complex linguistic situations, thereby learning correct values and language expression methods.[24]

4.3 Model Evaluation and Monitoring

4.3.1 Establishing an Ethical Evaluation Indicator System

In addition to traditional translation quality assessment indicators, it is imperative to construct a



specialized evaluation indicator system targeting ethical issues. The degree of bias can be measured by analyzing the model's output for descriptive tendencies towards different groups. Using sentiment analysis and semantic mining tools, one can determine whether the model exhibits inconsistent praise or criticism or stereotypical expressions when translating texts involving different races, genders, ages, and other groups.[25] Privacy leakage risks can be assessed based on security vulnerabilities and the likelihood of information leakage during data processing, such as detecting encryption flaws in data storage, transmission, and usage. The authenticity of information can be judged by comparing it with authoritative data sources. For translated news, academic materials, and other content, comparisons should be made with the original texts released by official sources or authoritative databases to ensure the truthfulness and accuracy of the translations. Through these quantifiable indicators, a comprehensive and objective evaluation of the model's ethical performance can be achieved.

4.3.2 Continuous Monitoring and Feedback

Establishing a continuous monitoring mechanism in the actual application of models is crucial for ensuring that models meet ethical requirements. Real-time collection of user feedback and usage data, combined with big data analytics, allows for dynamic assessment of model outputs. For example, by collecting user evaluations and questions about translation results through online translation platforms, high-frequency issues and anomalies in user feedback can be analyzed to promptly identify ethical problems that arise during model application. Once problems are detected, a swift adjustment and optimization process should be initiated. Depending on the severity and type of the issue, corresponding measures such as retraining parts of the model or adjusting parameter settings should be taken to ensure that the model always complies with ethical requirements.

4.3.3 Third-Party Auditing and Evaluation

Regularly inviting third-party organizations to conduct ethical audits and evaluations of models can effectively ensure the objectivity and fairness of the assessments. Third-party organizations typically possess specialized knowledge and an independent perspective, enabling them to conduct in-depth reviews of models from various dimensions.[26] For example, professional ethics research institutions can evaluate models from the perspectives of social and cultural ethics, while security testing organizations can conduct comprehensive inspections of data

security and privacy protection measures. After evaluation, third-party organizations will provide targeted and constructive opinions and suggestions, which are of significant reference value for model developers to better identify and address ethical issues.

4.4 User Education and Communication

4.4.1 User Guidelines and Instructions

It is necessary to provide users with detailed guidelines and instructions. The guidelines should clearly inform users of the potential ethical issues and risks associated with the translation process of large language models, such as cultural biases and privacy leakage risks in translation results. At the same time, users should be guided on the correct use of models in an easy-to-understand manner. For example, users should be reminded to use the model cautiously when translating texts involving personal privacy or sensitive information and to take necessary protective measures, such as anonymizing sensitive information before translation. Actual cases can be used to demonstrate the ethical issues that may arise from improper use of the model, thereby enhancing users' risk awareness.

4.4.2 Feedback Channel Construction

Establishing convenient user feedback channels and encouraging users to actively report ethical issues with the model are essential. Dedicated customer service email addresses, online feedback forms, or social media feedback portals should be set up, with professional customer service personnel responsible for handling user feedback. Customer service personnel should promptly respond to user inquiries and meticulously record and categorize feedback related to ethical issues. After regularly analyzing and summarizing the feedback, the information should be relayed to the model development team to provide important references for model optimization. For example, if users report misunderstandings in the translation of specific cultural texts, the development team can optimize the model accordingly.

4.4.3 Ethical Education and Publicity

Conducting ethical education and publicity through multiple channels, including official websites, social media, and offline activities, can help raise users' awareness and understanding of ethical issues in translation. Publishing popular science articles and case analyses on translation ethics on the official website, in an illustrated format, can introduce ethical guidelines and common issues in translation to users. Interactive discussions on social media platforms, with dedicated topics, can attract user participation



and allow them to share their experiences and feelings when using translation tools. Hosting offline lectures and training sessions, inviting experts and scholars to explain translation ethics, can enhance users' ethical awareness and sense of responsibility, promoting the healthy application of large language models in translation.

V. Conclusion

The widespread application of large language models in the field of translation has brought unprecedented convenience to global communication, greatly facilitating the dissemination of information and the sharing of knowledge. However, this study clearly reveals the many ethical issues lurking behind these models, which cover key dimensions such as accuracy and fidelity, the dissemination of culture and values, and user privacy and data security. These issues have a significant impact on translation quality, cross-cultural communication, and user rights.

In terms of accuracy and fidelity, large language models pose risks of information loss, incorrect translation, and distortion of source language culture and author intent. In professional fields such as medicine, law, and academic research, these issues can lead to severe consequences, mislead public perceptions, and hinder academic development. In the realm of cultural and value dissemination, cultural hegemony, the spread of cultural biases, and inappropriate influence on target language cultures disrupt equal cultural exchange and the coexistence of diverse cultures, potentially leading to misunderstandings and conflicts. User privacy and data security issues are particularly alarming, as data breaches not only infringe upon individual rights but also lead to crises of trust.

To address these issues, this study systematically proposes a range of mitigation strategies across four key areas: data processing, model design and training, model evaluation and monitoring, and user education and communication. In the data processing stage, diversified data collection, rigorous data cleaning and annotation review, and enhanced privacy protection provide a high-quality and secure data foundation for model training. During model design and training, the introduction of ethical constraint mechanisms, iterative optimization, and timely human intervention and guidance help avoid ethical issues from the outset and improve the model's ethical performance. Establishing a comprehensive ethical evaluation indicator system, continuous monitoring and feedback mechanisms, and third-party auditing and evaluation ensures that models meet ethical requirements during application. Providing detailed user guidelines and instructions, building

convenient feedback channels, and conducting extensive ethical education and publicity enhance users' risk awareness and ethical responsibility, promoting the healthy application of large language models.

This study not only enriches the theoretical research on the ethics of large language models in translation, constructing a systematic ethical analysis framework and laying a solid foundation for future research, but also promotes interdisciplinary integration. In practice, it provides valuable guidance for the translation industry and social development, helping to improve translation quality and promote fair, harmonious, and stable cross-cultural communication.

However, research on the ethical issues of large language models in translation is still in a stage of continuous development. In the future, as technology continues to advance and application scenarios expand, new ethical challenges will inevitably emerge. On one hand, it is necessary to further investigate the internal mechanisms of large language models and enhance their interpretability to better understand and control model behavior and fundamentally address ethical issues. On the other hand, continuous attention should be paid to ethical differences across different cultural backgrounds, and more universal yet flexible ethical guidelines should be developed to meet the needs of diverse cultural exchanges in a globalized context. Moreover, international cooperation and communication should be strengthened to jointly address the global ethical challenges posed by large language models in translation, promoting the sustainable development of artificial intelligence technology in the field of translation and ensuring that it better serves human society.

References

- [1]. Xu Yuemei et al., Technical Application Prospects and Risk Challenges of Large Language Models. *Computer Applications*, 2024. 44(06): pp. 1655-1662.
- [2]. Yu Tianbo, A Study on the Translation Features of ChatGPT's Chinese-English Translation of the 2023 Government Work Report. *Language and Culture Studies*, 2024. 32(04): pp. 154-158.
- [3]. Feng Zhiwei & Zhang Dengke, Large Language Models in Artificial Intelligence. *Foreign Languages*, 2024. 40(3): pp. 1-29.
- [4]. Mou Yiyang, Chen Hanxiao & Li Hongwei, Research Progress on Security and Privacy Protection Technologies for Large Language Models. *Journal of Cyberspace Security Science*, 2024. 2(1): pp. 40-49.



- [5]. Zhang Jinyi, Guo Cong & Gao Zhonghui, Research Progress on Neural Machine Translation Based on Linguistic Knowledge. *Artificial Intelligence and Robotics Research*, 2023. 12(2): pp. 97-106.
- [6]. Zheng Yinghao et al., Research and Design of Machine Translation Based on Pre-trained Models. *Science and Technology Innovation*, 2023(21): pp. 34-37.
- [7]. Wei Youwu, Li Na & Zhao Liangwei, A Study on Machine Translation Quality, Frequent Error Types, and Solutions: Based on the History of Machine Translation. *Modern Linguistics*, 2022. 10(9): pp. 1944-1949.
- [8]. Liao Shuyan, A Review of Data Cleaning Research. *Computer Knowledge and Technology*, 2020. 16(20): pp. 44-47.
- [9]. Li Zheng, Back to Ethics—The Future Path of Translation Ethics Research. *Journal of Zhejiang University of Commerce*, 2023(01): pp. 24-32.
- [10]. Shao Fengming & Zhai Changhong, Reinterpreting the "Fidelity" of Translation from the Perspective of Translator Behavior. *Journal of Qiqihar Normal University*, 2021(04): pp. 78-80.
- [11]. Zhao Jiu Xiao, On "Creative Treason" in Literary Translation from the Perspective of Translation Ethics. *Modern Linguistics*, 2023. 11(3): pp. 906-911.
- [12]. Guo Xiaodong, Risks of Generative Artificial Intelligence and Inclusive Legal Governance. *Journal of Beijing Institute of Technology (Social Sciences Edition)*, 2023. 25(06): pp. 93-105+117.
- [13]. Xu Yuemei et al., Technical Application Prospects and Risk Challenges of Large Language Models. *Computer Applications*, 2024. 44(6): pp. 1655-1662.
- [14]. Jiang Xueying & Xu Jing, Generative AI News in Human-Computer Interaction: Empowerment, Crises, and Responses. *Henan Social Sciences*, 2023. 31(12): pp. 105-113.
- [15]. Wu Lan, A Study on Enhancing English Intercultural Communication Skills in the Context of Artificial Intelligence. *Journal of Liaoning Open University*, 2023(02): pp. 72-75.
- [16]. Sun Fei & Wang Xiaochen, On the Translation Methods and Issues of Pearl S. Buck's Translation of Water Margin. *Modern Linguistics*, 2021. 9(4): pp. 942-946.
- [17]. Xu Yuemei et al., Research Progress and Implications of Large Language Models and Multilingual Intelligence. *Computer Applications*, 2023. 43(S2): pp. 1-8.
- [18]. Xu Qi, Empowering the Upgrading of Omnimedia Communication Infrastructure and Innovation of Application Ecosystem with Artificial Intelligence Large Models. *Publishing Horizon*, 2024(03): pp. 13-20.
- [19]. Zhu Yanlin, Innovations and Applications of Data Journalism in the Context of Intelligent Media. *Communication Power Research*, 2023. 7(20): pp. 55-57.
- [20]. Cui Shuang, Data Annotators: The Human Power Behind Artificial Intelligence. *Science Chinese*, 2019(19): pp. 72-73.
- [21]. Mou Yiyang, Chen Hanxiao & Li Hongwei, Research Progress on Security and Privacy Protection Technologies for Large Language Models. *Journal of Cyberspace Security Science*, 2024. 2(1): pp. 40-49.
- [22]. Qiu Desheng & Luo Yihong, Ethical Review of Algorithms Integrated into Public Decision-Making. *Theoretical Exploration*, 2024(02): pp. 5-12.
- [23]. Zhao Ruizhuo et al., Research Progress on Evaluation Technologies for Large Language Models. *Data Acquisition and Processing*, 2024. 39(03): pp. 502-523.
- [24]. Li Haili & Li Haoling, Foreign Publicity Translation of National Cultural Load Words in the Context of Artificial Intelligence. *Overseas English*, 2022(22): pp. 36-38.
- [25]. Zhou Changle, Intelligent Humanities Research and Its Future Development. *People's Tribune—Academic Frontier*, 2024(02): pp. 75-83+107.
- [26]. Xu Dan, Theories and Practices of Third-Party Evaluation in China. *Today's Science and Technology*, 2022(12): pp. 8-16+28.