ManCC: A Task-Anchored Benchmark for Manchu–Classical Chinese Cross-Lingual Modeling

Meiqi Wang[1], Xiaoxin Sun[2]*, Dongjie Wang[3]*, Ruixin Yu [2], Xiantao Heng [4], Shuo Wang [5], Zhen Huang [6], Peng Zhao [2], Suhua Wang[7]*, Minghao Yin[2]

[1] Jiangsu Second Normal University, [2] Northeast Normal University,
[3] University of Kansas, [4] Sichuan University, [5] Tonghua Normal University,
[6] University of Science and Technology of China, [7] Changchun Humanities and Sciences College

*Corresponding author: sunxx772@nenu.edu.cn, wangdongjie100@gmail.com, wangsuhua@ccrw.edu.cn

Abstract

Research in cross-lingual modeling for historical and extremely low-resource languages is hindered by the absence of standardized evaluation benchmarks. To address this, we present **ManCC**—the first task-anchored benchmark for **Man**chu–**C**lassical **C**hinese translation. ManCC consists of a high-quality parallel corpus of 16,627 sentence pairs, derived from the Qing-dynasty historical text Manwen Laodang-Taizong, and a reproducible evaluation protocol that combines automatic metrics (BLEU and chrF) with a three-dimensional human assessment (fidelity, fluency, linguistic normativity). Through systematic evaluation across three model families (non-pretrained, multilingual pretrained, and large language models), we find that linguistic differences significantly influence performance, broader language coverage in multilingual pretraining facilitates low-resource transfer, and automatic metrics often fail to capture essential errors in historical translation—underscoring the necessity of human evaluation. ManCC not only provides foundational resources for Manchu–Classical Chinese translation but also establishes a diagnosable, reproducible platform for cross-lingual modeling of historical low-resource languages [1].

1 Introduction

Mainstream research in natural language processing has predominantly focused on high-resource and modern languages, while critically endangered and historical languages have long lacked standardized, reproducible evaluation benchmarks (Joshi et al., 2020). The absence of such infrastructure has led to fragmented research efforts, making it difficult to compare methods horizontally or iteratively optimize findings, thereby severely hindering systematic progress in this direction. Manchu, as the "national language" of the Qing dynasty (Crossley and Rawski, 1993), carries nearly three centuries of precious historical archives but is now classified by UNESCO as **critically endangered** (Moseley, 2010), with only around fifty fluent readers and writers globally. The translation task between Manchu and Classical Chinese embodies multiple challenges of natural language processing under extremely low-resource scenarios: the two languages belong to agglutinative and isolating typologies, respectively, with vast typological differences; Classical Chinese exhibits a significant semantic gap with modern Chinese; moreover, domain experts with bilingual proficiency and deep knowledge of Qing dynasty texts are extremely scarce (Norman, 2003). These factors collectively result in an acute shortage of high-quality parallel corpora, and existing studies have mostly been small-scale exploratory cases (Zhang et al., 2024; Pei et al., 2025; Shu et al., 2024), which cannot support effective training or fair evaluation of deep learning models.

To address these core bottlenecks, the primary objective of this paper is not merely to improve machine performance for Manchu–Chinese translation but to position the translation task as an anchor task. Building upon the supervision signal this task provides, we aim to construct a reproducible evaluation benchmark for historical cross-lingual modeling. Through this systematic benchmark, we seek to diagnose the capabilities and limitations of different modeling paradigms on this challenging language pair, reveal systematic failure patterns driven by linguistic structural asymmetries, and thereby advance the field toward comparability and cumulative progress.

Specifically, the contributions of this paper are as follows:

- We present ManCC, a task-anchored benchmark for historical cross-lingual modeling, built upon a high-quality Manchu–Classical Chinese parallel corpus with expert-verified, sentence-level semantic alignment.

[1] https://zenodo.org/records/18149587

Findings of the Association for Computational Linguistics: ACL 2026, pages 27271–27292

- We introduce a reproducible evaluation protocol for extremely low-resource settings, combining standardized data splits, automatic metrics, and carefully designed human evaluation, which exposes key systematic limitations of commonly used translation metrics in distinct historical language scenarios.
- Through systematic benchmarking across diverse model families, we show that ManCC functions as a diagnostic testbed that meaningfully differentiates modeling paradigms and reveals recurrent failure modes driven by linguistic and structural asymmetries.

2 Related Work

2.1 Machine Translation for Low-Resource and Historical Languages

Machine translation for low-resource and historical languages presents a core challenge in cross-lingual processing, facing dual critical constraints: low-resource languages (e.g., Manchu) lack large-scale, high-quality bilingual parallel data, while historical languages (e.g., Classical Chinese) exhibit grammatical fossilization, semantic depth dependent on specific historical contexts, and insufficient standardization of proper nouns. Limited by both data scarcity and linguistic complexity, research in this area has long been marginalized in the NLP field. Existing studies primarily focus on two directions: data augmentation and model adaptation. At the data level, strategies such as back-translation Zebaze et al., 2025; Jia et al., 2025 and active learning Dossou et al., 2025; Guo et al., 2024 effectively alleviate data scarcity. At the model level, transfer learning via fine-tuning multilingual pre-trained models Louchheim et al., 2025 and quantized multilingual models Hb and Ptaszynski, 2025 has become a practical tool for low-resource translation. However, existing explorations generally suffer from severely limited data scale and highly inconsistent evaluation sets, making it difficult for truly fair method comparisons and reliable replication. Particularly for language pairs like Manchu–Classical Chinese, which embody both of these constraints, there remains a lack of sufficient parallel data support and a unified, rigorously designed standardized evaluation system.

2.2 Manchu Digitization and Text Recognition

The foundational digital work in Manchu studies has primarily concentrated on text recognition. Thanks to the construction of publicly available corpuss (e.g., the woodblock-printed Manchu word corpus WMW (Wang et al., 2022) and the handwritten corpus HMW (Wang et al., 2024)) and the application of deep learning models, Manchu OCR technology has made significant progress, achieving relatively high accuracy in word recognition (Zhang et al., 2021; Snowberger and Lee, 2024; Chung and Choi, 2025; Bi et al., 2025). These works provide crucial technical support for obtaining machine-readable text from original document images, serving as an important upstream step in the end-to-end Manchu NLP pipeline. However, their research scope is focused on character- or word-level visual recognition and has not yet involved core NLP tasks, such as sentence- or document-level semantic understanding and generation.

2.3 Evaluation Benchmarks and Methods

In the field of machine translation, automatic metrics based on n-gram overlap, such as BLEU (Papineni et al., 2002), have long been dominant. However, when faced with Manchu, with its complex morphological changes, and Classical Chinese, with its concise phrasing and flexible syntax, these metrics reveal clear shortcomings: they struggle to judge whether affixes are used correctly, whether historical function words are authentic, or whether culture-specific terms are accurately translated (Chen et al., 2025; ElNokrashy and Kocmi, 2023; Seo et al., 2023) To date, no publicly available evaluation benchmark with systematic human assessment dimensions has been established for the Manchu–Classical Chinese translation task (Chen et al., 2025; Nehrdich et al., 2025; Seo et al., 2023). This gap makes it difficult to quantify the practical value of models: first, there is no unified standard to determine whether translations conform to historical document writing norms and can directly serve academic research; second, it prevents the research community from forming a concerted effort on the core issue of "how to improve models' ability to capture historical language norms."

In summary, despite preliminary work on Manchu text recognition and scattered translation explorations, this field has consistently lacked an infrastructure centered on a sentence-level translation task, specifically incorporating high-quality parallel corpora, and equipped with a standardized evaluation protocol that effectively balances automatic and human assessment. The core purpose of our work is to systematically fill this gap.

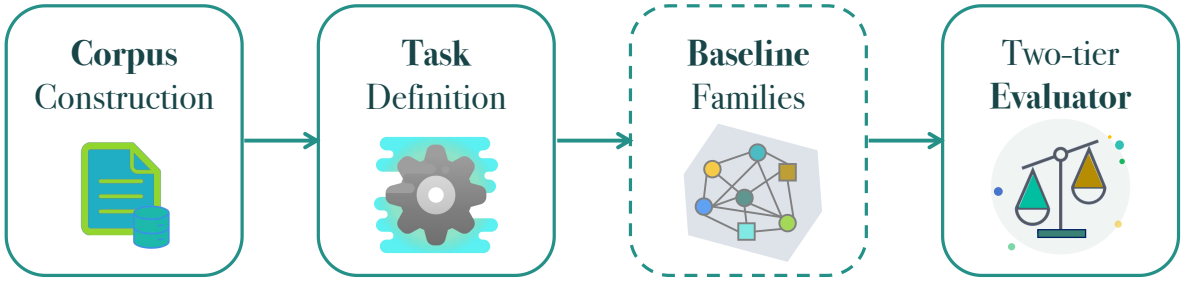

Figure 1: The ManCC benchmark aims to provide a standardized evaluation environment for the Manchu–Classical Chinese cross-lingual modeling scenario, which includes sentence level aligned corpus, a major two-direction translation task, reproducible baseline models, and automatic and manual evaluation indicators.

3 ManCC Benchmark Overview & Design

The ManCC benchmark aims to provide a well-structured, standardized and reproducible evaluation environment for the Manchu–Classical Chinese cross-lingual modeling scenario (Figure 1), which is characterized by extremely low-resource conditions and great linguistic difficulty. Its design adheres to the following core principles and consists of five carefully defined key components.

3.1 Design Principles

- Historical Fidelity Principle: The corpus is directly sourced from the original historical document Manwen Laodang-Taizong without any modern adaptation. This ensures that the benchmark task reflects truly genuine research challenges and that evaluation outcomes possess authentic historical validity.

- Sentence-Level Semantic Independence Principle: Through expert sentence segmentation and alignment, each parallel sentence pair is ensured to be semantically complete and independent. This provides a clearly defined unit for model training and evaluation, forming the foundation for reproducible benchmarking.

- Bidirectional Support Principle: The benchmark supports full bidirectional translation. This allows us to actively probe models for clear-cut asymmetric performance when processing different source language types (agglutinative vs. isolating), thereby diagnosing specific shortcomings in their cross-lingual transfer generalization capabilities.

- Evaluation Reliability Principle: It incorporates customized human evaluation metrics (Fidelity, Fluency, Linguistic Normativity) to compensate for the shortcomings of general-purpose automatic metrics (e.g., BLEU) in measuring the grammatical normativity of low-resource historical languages.

- Reproducibility Principle: The benchmark is released alongside open-source code and the corpus. This not only ensures direct comparability of results but also provides a clear framework for future expansion of the benchmark or adaptation to other historical languages.

3.2 Benchmark Components

The ManCC benchmark consists of the following four carefully designed, interconnected components, offering researchers a complete closed loop from high-quality data to standardized evaluation:

- Data Component: The core is a Manchu-Classical Chinese parallel corpus (size: 16,627 sentence pairs). The corpus is sourced from authoritative historical documents and constructed by a five-expert team through a three-stage process of independent translation $\rightarrow$ cross-validation $\rightarrow$ arbitration and finalization, ensuring high quality and consistency.

- Task Component: The main task is defined as Manchu-Classical Chinese bidirectional machine translation. This task provides explicit, clear supervision signals, and it can serve as a critical foundation for broader research, such as cross-lingual retrieval, historical linguistics, or computer-aided historical research.

- Evaluation Component: This comprises a two-tier evaluation protocol. (a) Automatic Evaluation Layer: BLEU and chrF. (b) Human Evaluation Layer: Introduces a human assessment

Indicator	Manchu (Char)	Manchu (Word)	Classical Chinese (Char)	Classical Chinese (Word)
Mean	61.97	13.02	16.69	9.49
Median	54.00	11.00	14.00	8.00
Min	3	1	1	1
Max	314	67	112	54
Std. Dev.	36.35	7.69	10.37	5.75
Total	1,030,359	9,583	277,453	28,414

Table 1: Statistics of metrics for the Manchu-Classical Chinese Corpus.

framework across three dimensions—Fidelity, Fluency, and Linguistic Normativity (using a 1–5 point scale). This component is crucial for diagnosing models' true performance in historical language scenarios.

- Baseline Component: It provides three categories of baseline models covering different modeling paradigms, including: untrained models (e.g., Transformer), pre-trained models (e.g., M2M100, NLLB), and large language models (e.g., DeepSeek-V3). These baselines offer a performance reference and a basis for comparison for subsequent research.

4 Dataset Construction

The core corpus of the ManCC corpus originates from the Manwen Laodang-Taizong, a Qing dynasty document of truly exceptionally high historical value. As there is no publicly available bilingual version of this text, we assembled an annotation team consisting of five experts (including three Manchu philology scholars and two Manchu-Classical Chinese bilingual translation experts) to construct a meticulously high-quality parallel corpus manually. The entire annotation process spanned several months and employed a rigorous three-stage quality control protocol:

- First, two experts independently performed the transcription of the original Manchu archival text and its precise preliminary translation into Classical Chinese.
- Subsequently, two other experts conducted sentence segmentation, cross-verification, and semantic alignment of the translations, ensuring that each sentence pair was semantically independent and accurately corresponded.
- Finally, a senior scholar acted as an arbiter, conducting the final comprehensive review, adjudication, and unified correction of any contested disputed sentence pairs.

mini gvniha bithe nikan han de isinara, 我以忠言致书于明帝， **nikan han -i bithe minde isinjire,** 明帝亦以书报我， **ishunde hafukiyame gisurefi doro aqaqi, akdun dere,** 彼此通达明晰后，则和好方牢固耳。 **gvniha gisun be gisureburakv,** 若不表衷情， **suweni gvniha de aqabume gisurefi doro aqaqi ombiu,** 顺从尔意，岂能和好。

Table 2: A bidirectional translation example (Manchu–Classical Chinese) from the ManCC parallel corpus. The corresponding English meaning is: I sent this letter to the Ming Emperor with sincere words, and the Ming Emperor also replied to me in writing. Only when our understandings were clear and mutual could our peace and amity be firmly established. If I did not lay bare my true intentions and simply yielded to your will, how could we achieve lasting peace?

This systematic workflow maximally guaranteed the high reliability of the corpus in terms of transcription accuracy, strict translation consistency, and sound semantic alignment quality.

Through the above process, we ultimately constructed the ManCC corpus comprising 16,627 high-quality parallel sentence pairs. Statistical analysis in Table 1 reveals that the corpus covers 9,583 distinct Manchu vocabulary items, fully reflecting the lexical diversity of the source material. The Classical Chinese portion exhibits an even richer vocabulary of 28,414 words, highlighting the characteristics of Classical Chinese such as "single-character words" and frequent use of synonyms. Table 2 presents a corpus sample, with its vertical Manchu script provided in Appendix F. The corpus topics encompass various types of Qing dynasty social activities, including politics, military affairs, bestowals, and daily administrative affairs, demonstrating good content diversity.

We randomly split the entire corpus into training

(13,301 sentence pairs), validation (1,662 sentence pairs), and test (1,664 sentence pairs) sets in an 8:1:1 ratio, using a fixed random seed to ensure consistency in data partitioning across different studies. More detailed annotation and analysis information of the corpus is presented in the supplementary Appendix A.

5 Tasks & Evaluation Protocol

5.1 Task Definition

The core task of this benchmark is bidirectional sentence-level translation between Manchu and Classical Chinese, namely Manchu → Classical Chinese and Classical Chinese → Manchu. This task is not only the direct objective for automating the translation of Manchu archives but also serves as a well-defined anchor task with clear supervision signals, providing a standardized test environment for measuring models' cross-lingual representation and generation capabilities in extremely low-resource, high-divergence scenarios.

5.2 Automatic Evaluation Metrics

To quantitatively assess translation quality, we adopt BLEU and chrF (Popović, 2015)—widely used metrics in machine translation—as core automatic indicators. BLEU measures fluency and common-word accuracy based on n-gram matches. However, considering the morphological complexity of Manchu as an agglutinative language and the unique tokenization characteristics of Classical Chinese, word-level metrics alone may not fully reflect alignment quality at morphological or character levels. Therefore, we additionally introduce the chrF, which is based on the F-score of character-level n-grams. This metric is more sensitive to rich morphological variations and can effectively compensate for the shortcomings of BLEU. When computing metrics for the Chinese side, we preprocess both the translation output and the reference using the jieba tokenizer to ensure evaluation consistency.

5.3 Human Evaluation Metrics

Given the inherent limitations of automatic metrics in capturing authentic historical language normativity and the translation accuracy of culture-specific terms, we designed a multi-faceted fine-grained human evaluation framework (more information is provided in Appendix B). This framework consists of three dimensions, each rated on a 1–5 scale (higher scores indicate better quality):

- Fidelity (Fide.): Assesses whether the translation completely and accurately reproduces the semantic information, factual details, and stylistic tone of the source text, without omission, addition, or distortion. We employ fidelity rather than the conventional adequacy, as historical translation requires accurate representation of cultural proper nouns, institutional terminology, and historical details, not merely information preservation.

- Fluency (Flue.): Assesses whether the translation is naturally smooth, idiomatic, and conforms to the core grammatical norms and standard expressive conventions of the target language (Classical Chinese or Manchu). We decouple linguistic normativity from fluency, since historical translation frequently produces results that are fluent but inconsistent with historical norms—for instance, rendering classical Chinese using modern Chinese grammar—thus demanding a dedicated evaluation criterion.

- Linguistic Normativity (Norm.): Tailored to the specificities of the target language, this dimension evaluates whether the translation complies with its historical grammatical, lexical, and orthographic norms (e.g., correct use of case particles and verb morphology in Manchu; appropriate usage of function words and sentence structures in Classical Chinese).

5.4 Evaluation Setup

For automatic evaluation, models will make predictions on the fixed test set (1,664 sentence pairs), and BLEU and chrF scores will be computed separately for each translation direction.

For human evaluation, we perform stratified sampling from the model predictions on the test set (200 samples per translation direction) to ensure coverage of samples from difficulty levels (sentence length). Each sampled instance will be independently and blindly rated by five domain experts proficient in both Manchu and Chinese, following the three-dimensional framework described above. Finally, we will report average scores and inter-rater agreement measures to ensure the statistical reliability of the human assessment.

Dir.	Type	Model	B-1↑	B-2↑	B-3↑	B-4↑	chrF↑	Fide.↑	Flue.↑	Norm.↑
Manchu ↓ Chinese	N-PLM	Transformer	31.05	18.93	13.01	9.48	17.42	2.85	3.17	3.02
	PLM	MarianMT	0.01	0.00	0.00	0.00	0.02	1.00	1.00	1.00
		Mbart-large-cc25	45.78	35.42	28.49	23.57	31.07	3.74	4.00	3.84
		Mbart-large-50	47.05	36.48	29.36	24.31	31.83	3.85	4.12	3.98
		M2M100-418M	46.45	35.25	27.77	22.53	30.96	3.97	4.22	4.10
		M2M100-1.2B	44.57	33.39	25.90	20.71	29.29	4.02	4.24	4.04
		Nllb-200-600M	42.83	31.44	24.08	19.03	27.94	3.54	3.83	3.71
		Nllb-200-1.3B	43.41	32.02	24.50	19.32	28.53	3.90	4.13	4.03
	LLM	Qwen	46.58	34.57	28.06	23.67	29.29	3.19	3.52	3.40
		DeepSeek	56.65	45.21	38.11	33.08	38.25	3.95	4.21	4.10
Chinese ↓ Manchu	N-PLM	Transformer	37.28	25.73	19.04	14.67	41.46	2.71	3.44	3.49
	PLM	MarianMT	62.78	52.17	44.51	38.75	63.26	3.82	4.00	4.17
		Mbart-large-cc25	62.59	52.29	44.88	39.25	63.01	3.83	4.03	4.15
		Mbart-large-50	63.90	53.71	46.24	40.54	63.91	3.90	4.04	4.22
		M2M100-418M	64.67	54.34	46.84	41.06	64.61	4.09	4.26	4.44
		M2M100-1.2B	60.89	50.15	42.32	36.47	62.56	4.14	4.21	4.27
		Nllb-200-600M	61.83	51.00	43.14	37.17	63.02	3.75	3.94	4.12
		Nllb-200-1.3B	62.71	51.84	44.04	38.11	63.10	4.07	4.25	4.39
	LLM	Qwen	36.52	24.94	17.15	11.95	57.32	3.50	3.77	4.00
		DeepSeek	62.44	51.83	44.10	38.20	63.47	4.04	4.17	4.17

Table 3: Bidirectional translation results for Manchu-Classical Chinese. B-1 to B-4 denote BLEU-1 to BLEU-4. Higher metric values indicate better performance, with underlined figures denoting the column's highest score.

6 Experiments Setup

6.1 Baseline Models

To comprehensively assess the capabilities and limitations of different modeling paradigms on the ManCC task, we systematically compare three representative categories of baseline models:

- Non-pre-trained Models: The standard Transformer (Vaswani et al., 2017) (encoder-decoder architecture) is selected as the baseline to evaluate the learning ability of models starting entirely from scratch, using only the ManCC training data.
- Pre-trained Models: We select translation models pre-trained on multilingual data, aiming to examine the transferability of their cross-lingual knowledge to the low-resource Manchu–Classical Chinese pair. Specific models include: MarianMT (Junczys-Dowmunt et al., 2018)(lightweight and efficient), mBART-large-cc25 and mBART-large-50 (Liu et al., 2020) (multilingual denoising pre-training), NLLB-200-distilled-600M and NLLB-200-1.3B (Team et al., 2022) (covering 200 languages), M2M100-418M and M2M100-1.2B (Fan et al., 2021) (direct translation across 100 languages).
- Large Language Models (LLMs): DeepSeek-V3 (Liu et al., 2024) and Qwen-Plus (Yang et al., 2025) are selected as representatives of large-parameter models to explore their ability to handle this translation task without parameter updates via in-context learning.

6.2 Training and Inference Protocols

- Fine-tuning: For the Transformer and all pre-trained models, we perform full-parameter fine-tuning on the ManCC training set.
- In-Context Learning (ICL): For LLMs, we adopt the ICL paradigm. Using different sentence embedding (e.g., xlm-roberta-large (Ruder et al., 2019), bge-large-zh-v1.5 (Xiao et al., 2024)) and different shot count.

Detailed hyperparameter configurations and specific settings for fine-tuning and ICL retrieval are provided in Appendix C. All experiments are conducted based on fixed data splits and random seeds to ensure comparability and reproducibility.

7 Experiment Results and Analysis

7.1 Overall Performance Comparison

Table 3 summarizes the overall performance of the three model families on the ManCC test set. The results reveal a clear performance hierarchy and notable asymmetry between translation directions. Pre-trained models significantly outperform

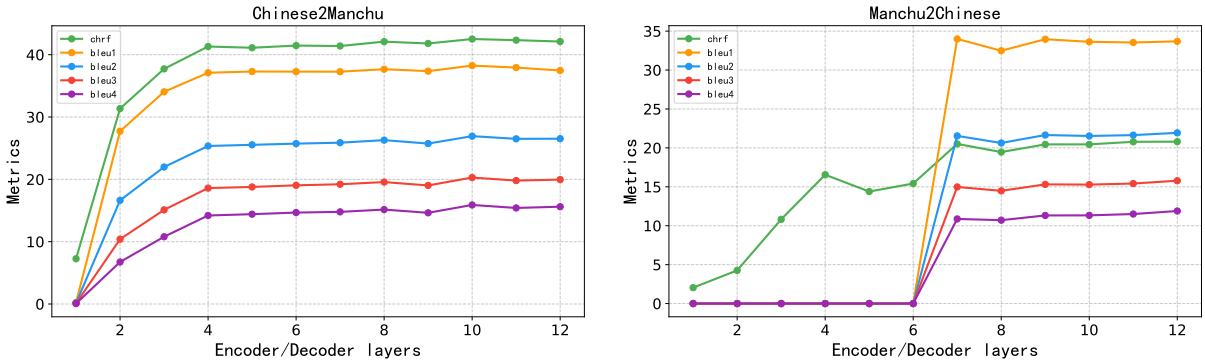

Figure 2: Performance trend of the Transformer model with varying model depth in Chinese-Manchu translation.

the untrained Transformer across automatic metrics (BLEU, chrF) and human evaluation, where inter-rater agreement achieves a Fleiss' Kappa of 0.78, highlighting the value of prior multilingual knowledge for this low-resource task.Untrained models exhibit very low automatic scores but can still learn basic semantic correspondences, as reflected in non-trivial human evaluation ratings. In contrast, LLMs operating purely via in-context learning achieve remarkably competitive performance, especially in the more challenging Manchu→Classical Chinese direction. This highlights their strong cross-lingual alignment and generation capabilities without any task-specific parameter updates. Human evaluations consistently provide a more nuanced assessment than automatic metrics, often capturing qualitative improvements in fidelity and normativity that are missed by surface-level n-gram matching. Additional prediction outputs of the various models are presented in Appendix D.

Interestingly, most models perform better when translating from Classical Chinese to Manchu than the reverse direction. This phenomenon can likely be attributed to four linguistic factors:(1) Writing system: Manchu is represented in Latin transliteration, which shares a subword space with multilingual pre-training corpora and facilitates knowledge transfer in the Chinese→Manchu direction. Classical Chinese uses logographic characters, requiring models to perform character-level semantic understanding and lacking direct alignment with Latin scripts, which significantly increases the difficulty of Manchu→Chinese translation. (2) Vocabulary size: The number of word types in Classical Chinese (28,414) is approximately three times that of Manchu (9,583). Manchu→Chinese translation requires precise lexical selection from a larger target vocabulary, leading to a higher risk of lexical errors. (3) Morphological type: Manchu is an agglutinative language that expresses grammatical categories such as case, number, tense, aspect, and voice via suffixes. Classical Chinese is an isolating language with no morphological inflection, relying mainly on function words and word order to encode grammatical relations. Manchu→Chinese translation focuses on morphological parsing and structural restructuring, while Chinese→Manchu translation emphasizes morphological generation; this asymmetry directly leads to divergent translation performance. (4) Syntactic structure: Manchu follows SOV word order, and case markers allow flexible constituent ordering. Classical Chinese is predominantly SVO, where grammatical relations are determined by surface position and discourse is highly paratactic. Manchu→Chinese translation must restructure flexible SOV constructions into position-sensitive SVO paratactic sentences, making word-order errors highly frequent.

7.2 Translation Direction Differences and Model Depth Requirements

To investigate how model architectures respond to varying source language complexity, we analyzed the performance variation of the untrained Transformer model as a function of encoder-decoder layer count (as shown in Figure 2). Results show that translating from morphologically complex Manchu requires substantially greater model depth (≥ 7) to achieve meaningful, stable performance, whereas translating from syntactically concise Classical Chinese saturates with far fewer layers (≤ 4). To investigate why shallow models underperform in Manchu→Chinese translation, we conduct a truncation experiment by limiting Manchu sentences to 25 words. Results show the bifurcation point drops from Layer 7 to 6. This suggests two sources

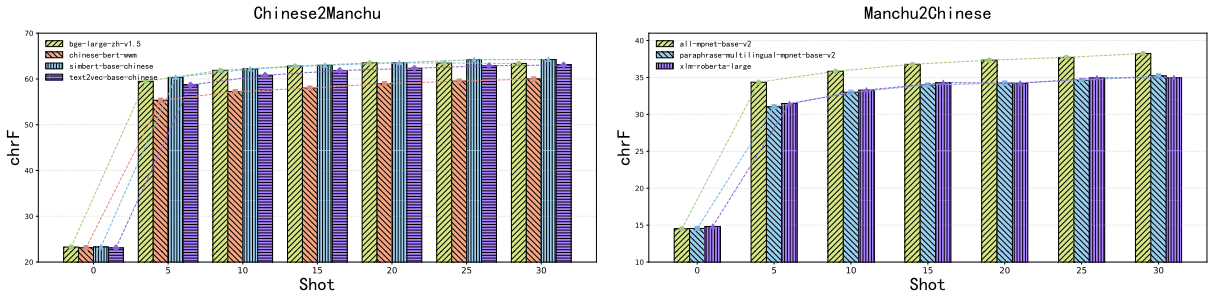

Figure 3: Impact of different shot counts and embedding models on the Manchu-Classical Chinese task.

of depth demand: morphological complexity requires at least 6 layers to handle case markers and suffixes, while long-distance dependencies in long sentences further raise the requirement to 7 layers. As shown in Figure 5, Manchu sentence length follows a long-tailed distribution, with many sentences containing complex cross-word dependencies. This explains why deeper models are needed in Manchu→Chinese translation.

7.3 Transfer Effects of Multilingual Pre-training

We examine the transferability of multilingual pre-training by comparing models with different language coverage. Models pre-trained on more languages consistently outperform those with narrower coverage, demonstrating that broader cross-lingual representations enhance adaptability to unseen low-resource languages like Manchu. Furthermore, performance does not scale monotonically with model size, indicating that successful transfer depends on the alignment between pre-trained knowledge, model architecture, and task-specific requirements, rather than parameter count alone.

7.4 Impact of Shot Count and Embedding Model in ICL

Figure 3 illustrates the impact of different shot counts and embedding models on the chrF score of DeepSeek in the bidirectional Manchu-Classical Chinese translation task. The results show that performance generally improves with an increasing number of shots, but with clear and consistent diminishing marginal returns, and the optimal embedding model type varies significantly across distinct translation directions.

7.5 Conflicts Between Automatic and Human Evaluation

The human evaluation and automatic metrics exhibit largely consistent trends across the majority of cases. However, in-depth analysis reveals critical and systematic misalignments, highlighting the limitations of over-reliance on automatic metrics in complex historical, low-resource language translation scenarios. We dissect the following two distinctly typical misalignment patterns:

For Case 1 (first line in Table 4), the core content word "singgeri" (rat), representing the "Year of the Rat" in the Chinese zodiac system, was incorrectly translated as the genitive particle "-i", completely altering the temporal reference—a major factual error. However, because the rest of the sentence “擒奸细一名” ("captured one spy") was translated correctly, this output still achieved a BLEU-1 score of 80.0 and a BLEU-4 score of 66.87, with automatic metrics indicating "excellent" performance. In human evaluation, however, its Fidelity score was only 2 out of 5, as the key information was distorted. This reveals BLEU’s reliance on surface n-gram matches and its inability to penalize severe mistranslations of core cultural/historical terms. Furthermore, while this translation scores low under traditional adequacy due to incomplete information, it performs even worse under our proposed fidelity metric, as it also distorts cultural proper nouns such as historical sexagenary cycle terms. This shows that historical translation requires a stricter fidelity standard.

Case 2 (second line in Table 4) is an exemplary case of high-quality paraphrase. While faithfully preserving all core semantics of the original, the model flexibly adapts the sentence structure to better align with the narrative conventions of historical Chinese. For instance, it concretizes “取册陈之” as “将封册自案上取下，置于东侧案上” and expands “概加宣读” to “概行宣读.” Such restruc-

Source	Target	Model Prediction
子年擒奸细一名。(In the Year of the Rat, a spy was captured.)	singgeri aniya emu giyansi jafaha,	-i aniya emu giyansi jafaha,
bithei hafan fungnere qe be dere qi gaifi, dergi ashan -i dere de sindafi, wasihvn forome ilifi qe de araha manju monggo nikan ilan gurun -i gisun be gemu hvlaha	文官取册陈之于东侧案上，西向立，其册载满蒙汉三体册文，概加宣读。(A civil official brought the register and placed it on the eastern table, then stood facing west. The register was filled with Manchu, Mongol, and Chinese texts, and he read out all its contents in full.)	文官将封册自案上取下，置于东侧案上，西向立，将册文内所载满洲、蒙古、汉三体文字概行宣读。(The civil official then took the sealed register from the table, placed it on the eastern table, and stood facing west. He read out all the Manchu, Mongol, and Chinese characters recorded in the register.)

Table 4: Two typical cases of Manchu-Classical Chinese translation. English translations are given in parentheses following the Chinese sentences.

turing enhances the fluency and historical vividness of the translation but significantly reduces n-gram overlap with the reference, resulting in a low BLEU-4 score of 16.33. Human evaluation, however, awards perfect scores of 5 across all three dimensions—judging it not only as a factually correct translation but as an excellent rendering of historical text. This demonstrates that for languages like Manchu, which feature relatively loose sentence structures and rich verbal morphology, high-quality translation into Chinese often necessitates information integration and syntactic reconstruction—a process severely underestimated by literal-alignment-based metrics like BLEU. In addition, this case receives a high fluency score because the translation is grammatically correct and smooth in modern Chinese. Yet it obtains a low Linguistic Normativity score, as the wording is overly colloquial and inconsistent with the written norms of classical Chinese. This case illustrates that evaluating only fluency is insufficient to judge whether a translation conforms to the historical stylistic requirements in historical language translation.

These cases (three more cases in Appendix E) reveal systematic blind spots of automatic metrics like BLEU in historical translation: (1) insensitivity to mistranslations of core cultural/historical terms, (2) penalization of legitimate syntactic restructuring, (3) inadequate weighting of errors in critical function words or predicates, and (4) inability to judge semantic fidelity. Since these failure modes directly concern the fundamental quality of historical document translation, our fine-grained human evaluation framework is not merely supplementary but serves as a necessary corrective and an indispensable gold standard for evaluating translation in historical low-resource language settings.

8 Conclusion

This study introduces ManCC, the first standardized benchmark for historical low-resource translation between Manchu and Classical Chinese. We release a high-quality parallel corpus of 16,627 sentence pairs, carefully aligned and validated by domain experts, with a reproducible evaluation protocol combining automatic metrics and a three-dimensional human assessment framework. Experiments across model families reveal key insights: The linguistic differences between Manchu and Classical Chinese significantly influence model depth requirements. Multilingual pre-training shows strong transfer potential, with broader language coverage improving performance. Automatic metrics like BLEU often misalign with human judgments on semantic fidelity and authenticity in historical contexts. These findings emphasize the need for linguistically aware modeling and human evaluation in historical language processing. ManCC provides curated data and diagnostic evaluation, serving as a foundational resource for this language pair and a reproducible template for other endangered and historically situated languages.

9 Acknowledgments

This work was supported by: (1) National Social Science Foundation in 2024 (No. 24BMZ101); (2) 2023 Project of the 14th Five-Year Plan for Scientific Research of the State Language Commission (No. YB145-73); (3) Jilin Provincial Science and Technology Development Program Project(No. 20250203109SF). We gratefully acknowledge Northeast Normal University Manchu Intelligent Information Processing and Digital Qing History Laboratory for providing computing resources and support for this research.

Limitations

As an initial attempt to construct a benchmark for historical low-resource languages, this study has the following limitations: (1) Homogeneity of Data Source: The ManCC corpus is currently built entirely upon Manwen Laodang-Taizong. Although this document holds extremely high historical value and covers multiple themes such as politics, military, and social affairs, its single-source nature inevitably limits the textual genre diversity of the corpus (e.g., lacking other genres such as personal letters, literary works, or legal documents). This may affect models' generalization capabilities to other types of Manchu archives. (2) Scalability Bottleneck of Expert Annotation: The high quality of the corpus heavily relies on scarce Manchu-Classical Chinese bilingual experts for multiple rounds of manual annotation and validation. This process is time-consuming, labor-intensive, and costly, posing significant challenges for rapidly expanding the corpus scale or replicating it for other historical language pairs under the same standards, thereby constraining the benchmark's evolution speed and broader application. (3) Potential Subjectivity in Evaluation Dimensions: Although we designed a detailed human evaluation framework and employed multiple expert ratings to reduce subjective bias, judgments on aspects such as "linguistic normativity" and "historical style" still depend to some extent on the individual knowledge and experience of the evaluators. While we report inter-rater agreement, quantifying the "authenticity" of historical language in a fully objective manner remains an open challenge.

We explicitly acknowledge these limitations and believe they indicate important directions for future improvement and exploration.

10 Ethics/Data Statement

- Data Copyright and Source: The core source text of the ManCC benchmark, Manwen Laodang-Taizong, is a historical archival document from China's Qing dynasty, and its original text is in the public domain. The structured parallel corpus (ManCC) formed through our digital transcription, translation, alignment, and annotation of the original text will be publicly released by the author team in compliance with academic norms for non-commercial academic research use. Users must cite this paper and acknowledge the data source in related work.
- Cultural Sensitivity Statement: The texts processed in this study are official historical records from the Qing dynasty, primarily involving objective accounts of political, military, and social affairs. They do not contain negative evaluations of modern individuals, ethnic groups, or societies, nor do they involve sensitive content such as religion or privacy. We recognize that historical texts inherently carry perspectives and expressions specific to their era. Researchers using this data should adopt an objective and rigorous academic attitude, understanding its historical context.
- Annotation Work Statement: The construction of the ManCC corpus was underpinned by professional manual annotation conducted by a team of five specialists in Manchu studies. All annotators were recruited via professional associations dedicated to Manchu studies; compensation was aligned with both their expertise level and local academic consulting standards, and deemed fair and appropriate given the specialists' qualifications and the project's research budget constraints. This annotation work strictly adheres to academic ethics, involves no inappropriate content handling, and incurs no harm to any individuals or groups throughout the process.
- Potential Impact: We hope that the release of the ManCC benchmark will actively promote endangered language preservation, historical document digitization, and the development of cross-lingual artificial intelligence technologies. At the same time, we encourage users of this data to consider the potential impact of technological applications on historical interpretation and to strive to steer such applications toward responsible directions that enhance academic research capabilities and aid cultural heritage preservation.

References

Xiaojun Bi, Wenhao Tao, Zheng Chen, and Haipeng Sun. 2025. Scc3: a novel structure-connected cognition cube network for manchu word recognition. *Expert Systems with Applications*, page 129374.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min

Zhang. 2025. Benchmarking LLMs for translating classical Chinese poetry: Evaluating adequacy, fluency, and elegance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33007–33024, Suzhou, China. Association for Computational Linguistics.

Yan Hon Michael Chung and Donghyeok Choi. 2025. Finetuning vision-language models as ocr systems for low-resource languages: A case study of manchu. *arXiv preprint arXiv:2507.06761*.

Pamela Kyle Crossley and Evelyn S. Rawski. 1993. A profile of the manchu language in ch'ing history. *Harvard Journal of Asiatic Studies*, 53:63.

Bonaventure F. P. Dossou, Ines Arous, and Jackie CK Cheung. 2025. Rethinking full finetuning from pretraining checkpoints in active learning for African languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 64–78, Vienna, Austria. Association for Computational Linguistics.

Muhammad ElNokrashy and Tom Kocmi. 2023. eBLEU: Unexpectedly good machine translation evaluation using simple word embeddings. In *Proceedings of the Eighth Conference on Machine Translation*, pages 746–750, Singapore. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and 1 others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Jiaxin Guo, C. L. Philip Chen, Shuzhen Li, and Tong Zhang. 2024. DEUCE: Dual-diversity enhancement and uncertainty-awareness for cold-start active learning. *Transactions of the Association for Computational Linguistics*, 12:1736–1754.

Barathi Ganesh Hb and Michal Ptaszynski. 2025. RBG-AI: Benefits of multilingual language models for low-resource languages. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1233–1239, Suzhou, China. Association for Computational Linguistics.

Yepai Jia, Yatu Ji, Xiang Xue, Lei Shi, Qing-Dao-Er-Ji Ren, Nier Wu, Na Liu, Chen Zhao, and Fu Liu. 2025. A semantic uncertainty sampling strategy for back-translation in low-resources neural machine translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 528–538, Vienna, Austria. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, and 1 others. 2018. Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Carter Louchheim, Denis Sotnichenko, Yukina Yamaguchi, and Mark Hopkins. 2025. Using encipherment to isolate conditions for the successful fine-tuning of massively multilingual translation models. In *Proceedings of the Tenth Conference on Machine Translation*, pages 241–252, Suzhou, China. Association for Computational Linguistics.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

Sebastian Nehrdich, Avery Chen, Marcus Bingenheimer, Lu Huang, Rouying Tang, Xiang Wei, Leijie Zhu, and Kurt Keutzer. 2025. MITRA-zh-eval: Using a buddhist Chinese language evaluation dataset to assess machine translation and evaluation metrics. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 129–137, Albuquerque, USA. Association for Computational Linguistics.

Jerry Norman. 2003. The manchus and their language (presidential address). *Journal of the American Oriental Society*, 123(3):483–491.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schuetze. 2025. Understanding in-context machine translation for low-resource languages: A case study on Manchu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8767–8788, Vienna, Austria. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019. Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38.

Jean Seo, Sungjoo Byun, Minha Kang, and Sangah Lee. 2023. Mergen: The first Manchu-Korean machine translation model trained on augmented data. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 118–124, Singapore. Association for Computational Linguistics.

Peng Shu, Junhao Chen, Zheng Liu, Hui Wang, Zihao Wu, Tianyang Zhong, Yiwei Li, Huaqin Zhao, Hanqi Jiang, Yi Pan, Yifan Zhou, Constance Owl, Xiaoming Zhai, Ninghao Liu, Claudio Saunt, and Tianming Liu. 2024. Transcending language boundaries: Harnessing llms for low-resource language translation. *ArXiv*, abs/2411.11295.

Aaron Daniel Snowberger and Choong Ho Lee. 2024. Manchu script letters dataset creation and labeling. *Journal of Information & Communication Convergence Engineering*, 22(1).

NLLB Team, Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhiwei Wang, Siyang Lu, Mingquan Wang, Xiang Wei, and Yingjun Qi. 2022. Amre: An attention-based crnn fornbsp;manchu word recognition onnbsp;anbsp;woodblock-printed dataset. In *Neural Information Processing: 29th International Conference, ICONIP 2022, Virtual Event, November 22–26, 2022, Proceedings, Part II*, page 267–278, Berlin, Heidelberg. Springer-Verlag.

Zhiwei Wang, Siyang Lu, Xiang Wei, Run Su, Yingjun Qi, and Wei Lu. 2024. Learn more manchu words with a new visual-language framework. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025. TopXGen: Topic-diverse parallel data generation for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22358–22381, Suzhou, China. Association for Computational Linguistics.

Diandian Zhang, Yan Liu, Zhuowei Wang, and Depei Wang. 2021. Ocr with the deep cnn model for ligature script-based languages like manchu. *Scientific programming*, 2021(1):5520338.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand. Association for Computational Linguistics.

A Data Annotation Details and Statistical Analysis

A.1 Detailed Data Annotation Guidelines

Annotators were presented with the following disclaimer before starting the task: This task entails the quality assessment of historical text translations. All materials are excerpted from Qing-dynasty archival documents and contain no personally identifiable information (PII) or sensitive content. Your participation is entirely voluntary, and you may withdraw from the task at any time without penalty. All collected ratings will be exclusively used for academic research purposes, with all annotators-related data anonymized in subsequent publications.

All annotators are required to comply with the following annotation protocols, which govern the three-stage workflow as specified below:

- Transcription and Initial Translation: Convert the original Manchu text into Latin transliteration in strict accordance with the Roman Transcription Scheme, and provide a preliminary Classical Chinese translation that preserves the historical semantics of the source text.

- Sentence Segmentation and Alignment: Perform sentence-level segmentation on the translated text, ensuring that each Manchu–Classical Chinese sentence pair achieves strict one-to-one correspondence in

core semantic content. Appropriate syntactic adjustments are permitted to accommodate the typological differences between the two languages.

- Dispute Resolution: In cases of semantic ambiguity or unresolved translation discrepancies, annotators shall flag the relevant sentence pair as “pending arbitration” for final adjudication by a senior Manchu philology expert.

All annotation results are anonymized and will be used exclusively for academic research purposes. The payment was set at 0.04 RMB per annotated character. Prior to the annotation work, we fully negotiated and confirmed this compensation standard with all participating experts.

A.2 Visualization of Corpus Length Distribution

To visually illustrate the differences in sentence length between Manchu and Classical Chinese, we present histograms and box plots of character counts and word counts (Figure 4 and Figure 5). The charts clearly show that Classical Chinese characters are highly concentrated in the short-sentence range, while Manchu character distribution is more dispersed. At the word level, both exhibit right-skewed distributions, but Manchu shows a more pronounced long tail.

A.3 Word Cloud of Lexical Features

We generated word clouds for the high-frequency words (top 200) in Manchu and Classical Chinese (Figure 6). The Manchu word cloud centers on grammatical function words like "be" and "de," while the Classical Chinese word cloud highlights function words such as "之" and "其" as well as historical proper nouns like "贝勒" and "汗".

We employed the Latent Dirichlet Allocation (LDA) model for topic mining on the Classical Chinese text, identifying six topics. Figure 7 presents a heatmap of keywords under each topic, clearly outlining the semantic boundaries of different themes such as political interaction, material rewards, and military actions. Figure 8 shows box plots of the probability distribution of each topic across documents, reflecting the semantic diversity of the text and the stability of topic coverage.

These figures above visually present the stark contrast between the two languages in terms of core lexical function and textual focus, serving as a strong supplement to content feature analysis for Section 4.

B Detailed Human Evaluation Guidelines

Evaluators were presented with the following disclaimer before starting the task: This task entails the quality assessment of historical text translations. All materials are excerpted from Qing-dynasty archival documents and contain no personally identifiable information (PII) or sensitive content. Your participation is entirely voluntary, and you may withdraw from the task at any time without penalty. All collected ratings will be exclusively used for academic research purposes, with all evaluator-related data anonymized in subsequent publications.

Each evaluator is required to assess Manchu–Classical Chinese translation quality across three dimensions: faithfulness, fluency, and linguistic normativity. A 5-point Likert scale is adopted for each dimension, where 1 represents the lowest quality and 5 represents the highest. Evaluators must adhere to the detailed scoring criteria and operational definitions provided below, which also serve as supplementary specifications to Section 5.3.

B.1 Fidelity

- 5 points: Fully and accurately conveys all semantic information, factual details, and stylistic tone of the source text without any omission, addition, or distortion. Proper nouns and culture-loaded terms are translated precisely.
- 3 points: Conveys the core semantics but exhibits minor omissions of secondary information, slight ambiguity, or inaccurate translation of individual proper nouns, without altering the main point of the source.
- 1 point: Severely deviates from the source semantics, omits key information, or contains factual errors/fabricated content.

B.2 Fluency

- 5 points: The translation fully conforms to the grammatical norms and expressive conventions of the target language (Classical Chinese or Manchu), is coherent and idiomatic, and shows no translationese.
- 3 points: The translation is generally coherent but contains occasional awkward sen-

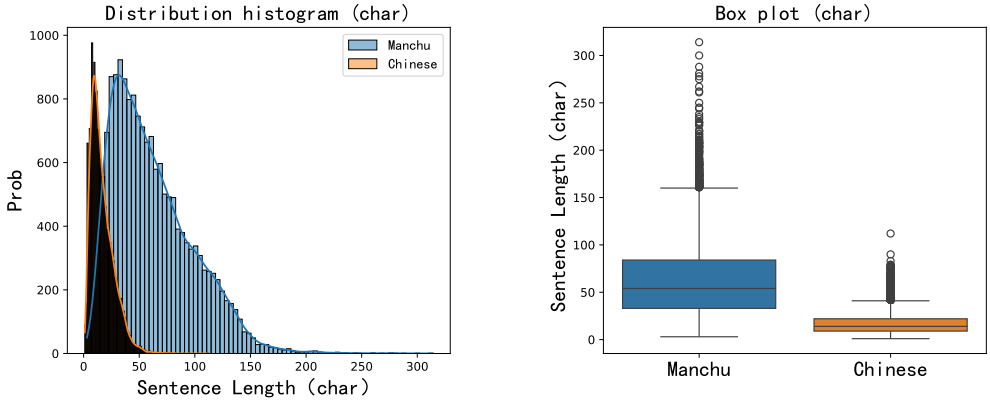

Figure 4: Histogram and box plot of text length (char count) distribution.

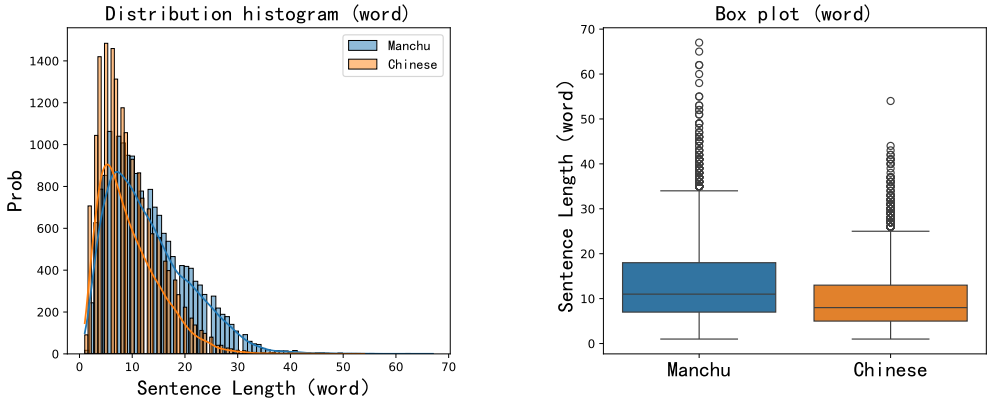

Figure 5: Histogram and box plot of text length (word count) distribution.

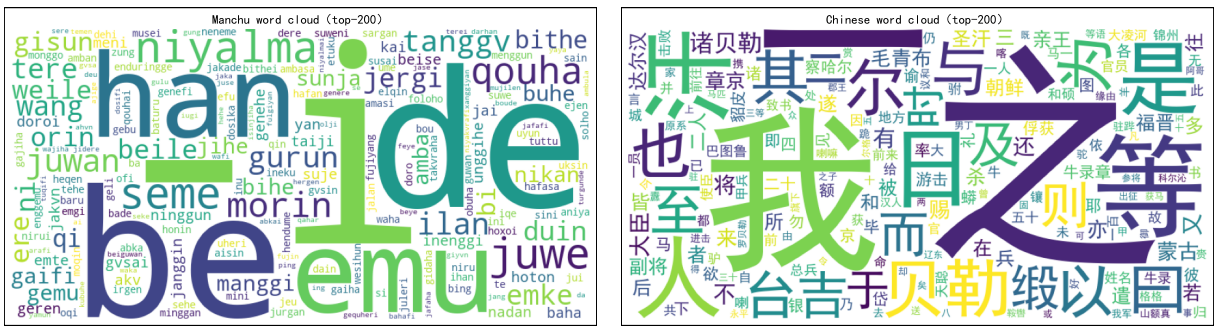

Figure 6: Word cloud of high-frequency Manchu and Chinese words.

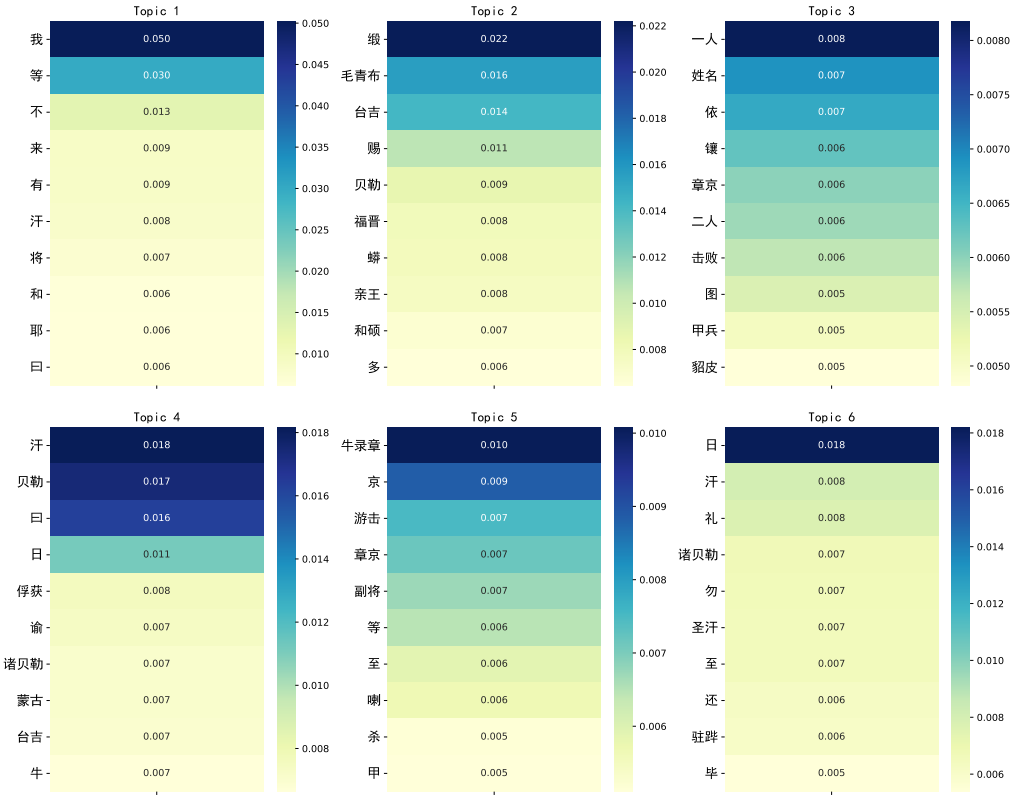

Figure 7: Heatmap of keywords for each topic.

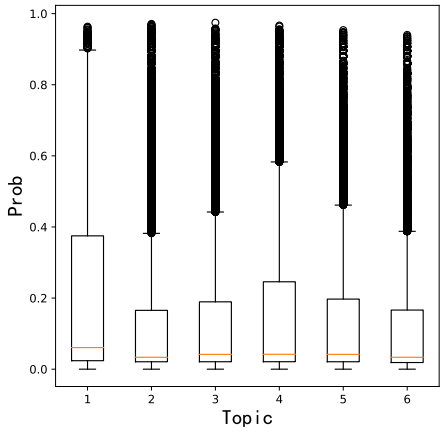

Figure 8: Box plot of probability distribution for each topic.

tence structures, inappropriate collocations, or somewhat unnatural expressions.

- 1 point: The translation is stiff and difficult to comprehend, contains numerous grammatical errors, and significantly hinders understanding.

B.3 Linguistic Normativity

- 5 points: The translation strictly adheres to the historical grammar and lexical norms of the target language. Manchu translations use suffixes and case particles correctly, with accurate verb morphology; Classical Chinese translations use function words and sentence structures in accordance with ancient Chinese norms, with careful word choice.

- 3 points: The translation is largely normative but contains occasional misuse of historical grammar, anachronistic vocabulary, or atypical yet comprehensible structures.

- 1 point: The translation severely violates historical language norms, exhibiting modern grammatical structures, misuse of function

Hyperparameter	Value/Option
Optimizer	AdamW
Learning Rate	{1e-5, 5e-5, 1e-4}
Training Epochs	{50, 100}
Batch Size	16
Maximum Sequence Length	128

Table 5: Hyper-parameter settings.

words, or coinage of vocabulary/forms that do not conform to historical usage.

C Detailed Experimental Configuration

This section provides the detailed experimental settings mentioned in Section 6.2. Hyperparameter selection was performed via grid search, where the final set of parameters was chosen based on the maximum validation performance. And we report the mean value across multiple independent runs as the final result.

C.1 Transformer Hyperparameters

We implement the Transformer as a standard encoder–decoder architecture based on the Hugging Face Transformers library. The basic hyperparameter settings are as follows: the hidden dimension is 768, the number of attention heads in both the encoder and decoder is 12, the feed-forward network dimension is 3072, the dropout rate is 0.1, and both encoder layerdrop and decoder layerdrop are set to 0.0.

C.2 Model Fine-tuning Hyperparameters

The Table 5 below summarizes the key hyperparameter settings shared by all models requiring fine-tuning (Transformer and pre-trained models). For MarianMT, we use the specific pretrained models from Hugging Face: Helsinki-NLP/opus-mt-en-zh for the Manchu→Chinese direction and Helsinki-NLP/opus-mt-zh-en for the Chinese→Manchu direction.

C.3 Details of ICL Settings

For the in-context learning (ICL) of large language models, we tested different configurations:

(1) Embedding Models for Retrieval:

- Manchu Sentence Embeddings: XLM-RoBERTa-large, paraphrase-multilingual-mpnet-base-v2, all-mpnet-base-v2.
- Chinese Sentence Embeddings: BGE-large-zh-v1.5, BERT-base-chinese, Chinese-BERT-wwm, text2vec-base-chinese.

(2) Number of In-context Examples (k): $k = 0, 5, 10, 15, 20, 25, 30$ shots. Retrieval is based on cosine similarity, selecting the k most similar sentence pairs from the training set as in-context examples for the input.

D Case Study

We present four prediction results from all models as illustrative cases, among which two are for Chinese → Manchu translation and the other two for Manchu → Chinese translation.

D.1 Chinese2Manchu-case-1

The case in Table 6 focuses on Qing imperial ritual scenarios, core evaluating historical context conformity and complex syntactic reconstruction. It emphasizes the translation of ritual-specific appellations such as "圣汗", "实录", "黄宝御案", and the coherence of the action chain "入-接-置", with significant model performance disparities:

M2M100-1.2B performs optimally (Fidelity: 4/5, Fluency: 5/5, Linguistic Normativity: 5/5), accurately reproducing standard Manchu translations of ritual terms (e.g., "实录" as "yargiyan kouli bithe") and conveying the ritual posture of "恭接" via "juwe galai dorolome".

Medium-performance models (Nllb-200-1.3B, M2M100-418M, Mbart-large-cc25, DeepSeek) have partial deviations, such as simplifying the documentary attributes of "实录" or omitting the spatial information of "由侧而入".

Low-performance models (Qwen, Nllb-200-600M, MarianMT, Transformer) perform poorly, with frequent translation errors of appellations or complete deviation from the original context.

In summary, in this case, only M2M100-1.2B fully meets ritual scenario translation needs. Medium and low-performance models struggle to align with Qing ritual contexts due to insufficient historical knowledge, highlighting the high demand for models' comprehensive capabilities in complex historical translation.

D.2 Chinese2Manchu-case-2

Table 7 focuses on Qing Manchu administrative rule texts, evaluating models' ability to reconstruct the three-tier logic of "scope delimitation - conditional judgment - rule definition" and adapt to

Model	Content
Source	圣汗由側而入，双手恭接实录，置于殿内黄宝御案上。(The Holy Khan entered from the side, respectfully received the Veritable Records with both hands, and placed them on the golden imperial desk inside the hall.)
Target	enduringge han ashan qi dosifi yargiyan kouli bithe be juwe galai dorolome alime gaifi, yamun -i dorgi suwayan bou iui dere de sindaha。
Transformer	enduringge han angga isire onggolo, ninggun mafa -i baru gisureme toktobuha manggi, iqi ergi ashan de wasihvn forome qaqari de dere qafi sarilaha
MarianMT	enduringge han ashan deri arame dosifi, juwe galai yargiyan arafi yamun -i dolo bou iui dere de sindahabi
Mbart-large-cc25	enduringge han, ashan -i dolo dosifi juwe galai alime gaifi, yamun -i dolo hvwang bou be dere -i ninggude sindahabi
Mbart-large-50	enduringge han hetu dosifi juwe galai wesihun录 be alime gaifi yamun -i dolo sindafi, boubai dere de sindaha
M2M100-418M	enduringge han ashan qi dosifi, juwe galai genggiyen yargiyan be okdome ganafi, yamun -i dorgiqi hvwang bou iui dere de sindaha
M2M100-1.2B	enduringge han ashan qi dosifi, juwe galai alime gaifi yargiyan kouli be yamun -i dolo hvwang bou iui dere -i dele sindambi
Nllb-200-600M	enduringge han ashan -i jihe be juwe galai dosifi, yargiyan kouli be juwe galai alime gaifi, yamun -i dolo suwayan bou iui dere de sindaha
Nllb-200-1.3B	enduringge han ashan -i dalba qi dosifi, juwe galai ging ni yargiyan be alime gaifi, yamun -i dorgi suwayan bou iui dere -i dele ilibuha
Qwen	enduringge han ashan de ilifi, juwe galai yargiyan kouli bithe alime gaifi, yamun -i suwayan bou iui dere de sindafi
DeepSeek	enduringge han ashan i duka ci dosifi, gala juwe i genggiyen i yargiyan kooli bithe be alime gaifi, diyan i dorgi suwayan booi iui dere de sindaha

Table 6: Classical Chinese → Manchu translation model predictions. English translations are given in parentheses following the Chinese sentences.

Model	Content
Source	所有编定品级之官员，其官职高者，依其官职之品级；(All officials with fixed ranks; those of higher official positions shall act in accordance with their ranks.)
Target	ere jergi banjibuha hafasa, ini hafan amba oqi, hafan -i jergi de oso,
Transformer	yargiyan qi toktobuha hafan -i hergen be jergi bodome xajin de alahabi,
MarianMT	jai jergi bodome tukiyehe hafan amban ton -i jergi, ini hafan -i jergi,
Mbart-large-cc25	yaya jergi be toktobuha hafasa, hafan hergen -i jergi,
Mbart-large-50	jing jergi bodome hafan oqi, hafan amban oqi, ini hafan -i jergi,
M2M100-418M	yaya jergi banjibuha hafasa de hafan amban oqi, ini hafan -i jergi,
M2M100-1.2B	yaya jergi banjibuha hafasa de hafan wesihun ujen niyalma, hafan -i jergi bodome ini hergen -i jergi,
Nllb-200-600M	yaya toktobuha hafan ini hafan -i jergi, hafan wesihun niyalma, ini hafan -i jergi,
Nllb-200-1.3B	ai jergi bodome sindaha hafasa, hafan amban oqi, ini hafan -i jergi bodome,
Qwen	jai jergi wesihengge hafasa de, hafan amban biqi, ini hafan -i jergi de etukini,
DeepSeek	jergi toktobuha hafasa de hafan amban oqi, ini hafan -i jergi,

Table 7: Classical Chinese → Manchu translation model predictions. English translations are given in parentheses following the Chinese sentences.

Manchu grammatical features (e.g., postpositive attributives, case markers). Performance disparities lie in semantic integrity, syntactic coherence and grammatical normativity.

M2M100-1.2B performs optimally (Fidelity: 5/5, Fluency: 5/5), fully reconstructing core logic: "yaya" accurately corresponds to the scope meaning of "所有", "oqi" conveys hypothetical relations, and "de" expresses "in accordance with". Its postpositive attributive structure fits Manchu grammar, with vocabulary and style consistent with Qing administrative document norms.

NLLB-200-1.3B (Fidelity: 4/5) retains core "condition - rule" logic but has semantic defects: "ai jergi" replaces "yaya jergi" (losing "所有"'s scope meaning) and lacks core verb "oso" (incomplete rule expression). M2M100-418M and DeepSeek (Fidelity: 3-4/5) show partial deviations (e.g., missing scope words) without breaking core logic.

Transformer (Fidelity: 1/5) completely deviates from original semantics/logic, only retaining "jergi" (rank). It mistranslates "编定品级" and "依其品级", loses key information like "所有" and "官职高者", with grammatical marker abuse and confused collocations, failing to fit Qing administrative document norms and historical context.

In summary, M2M100-1.2B excels in logical reconstruction and grammatical adaptation. Medium-performance models have minor semantic defects but maintain core logic, while Transformer fails entirely. Grasping administrative text logic and Manchu grammatical features is critical for high-quality historical text translation.

D.3 Manchu2Chinese-case-1

The case in Table 8 is a military operation order text from the Qing Dynasty, with the core objective of evaluating the models' translation accuracy regarding military narrative logic, action sequences and terminology. Significant performance disparities are observed among the models:

M2M100-1.2B and M2M100-418M deliver the optimal performance (Fidelity, Fluency and Linguistic Normativity: 5/5). Either of them accurately reconstructs the core action logic or fully presents the action chain of " 乘夜而入 - 暗中设梯 - 领众登城 - 攻克 ", which is consistent with the concise and coherent stylistic features of military orders.

Nllb-200-1.3B (Fidelity: 4/5) fails to include the core objective of " 攻克 ", leading to incomplete action description; DeepSeek (Fidelity: 3/5) incorrectly adds the irrelevant action of " 放火呐喊 ", deviating from the core requirement of " 暗中行动 ". Qwen and Transformer perform poorly (Fidelity: 2/5), either reversing the mode of the " 攻城 " operation or producing syntactically confusing expressions that do not conform to the norms of military orders.

The top-performing models accurately capture the Manchu military sentence structure characterized by "multi-action progression". Their translations of phrases such as "dobori dosifi" and "hvlhame wan sindafi" are consistent with the recording conventions in Veritable Records of the Qing Dynasty.

In summary, the M2M100 series stands out due to its precise deconstruction of military action sequences and narrative logic. The remaining models either have deviations in action details and omissions of objectives or present syntactic confusion, indicating that their capability in scenario adaptation and logical reconstruction for military texts still needs further improvement.

D.4 Manchu2Chinese-case-2

The case in Table 9 is a moral condemnation argumentative text from the Qing Dynasty, with the core objective of evaluating the models' translation accuracy in terms of abstract semantics, causal logic and solemn condemnatory stylistics. Significant performance disparities are observed across the models:

DeepSeek and Mbart-large-50 deliver the optimal performance (Fidelity, Fluency, Linguistic Normativity: 5/5). Their translated versions accurately reconstruct the causal logic of "贪利欺上 - 动机 - 恶行 - 后果" and adopt a solemn style that aligns with the text's requirements.

M2M100-1.2B and Nllb-200-1.3B achieve a Fidelity score of 3/5, with issues such as semantic deviations or omissions of core immoral acts; M2M100-418M and Nllb-200-600M score 2/5 in Fidelity, suffering from semantic confusion or rigid expressions; Qwen and Transformer perform the worst with a Fidelity score of 1/5, completely deviating from the original semantics.

In summary, models such as DeepSeek stand out for their accurate reconstruction of logic and semantics. The remaining models are afflicted by element omissions or semantic deviations, indicating that their capabilities in abstract text parsing and moral context adaptation need to be further enhanced.

Model	Content
Source	tereqi juleri tuqike ambasa, jakvnju niyalma uthai genefi dobori dosifi, hvlhame wan sindafi heqen be tafafi, tereqi musei geren qouha be gajifi tafambufi gaiha,
Target	先行之大臣及其所率八十人，由此速往，乘夜而入，暗中设梯，领我众军，登城竟之。(The leading ministers and the eighty men under their command hastened here, entered by night, secretly placed ladders against the wall, led our troops, scaled the city, and captured it.)
Transformer	我前遣大臣，即行诸至八十人路出＼诱我军士梯登城陷；(I previously sent ministers, leading eighty men out by the road to lure our soldiers to climb ladders and breach the city.)
MarianMT	to,,,,
mbart-large-cc25	于是,前队诸臣及八十人即遣夜入城,竖梯登城,率我兵及大军登城,克之。(Then the ministers of the vanguard and the eighty men were immediately sent to enter the city at night, set up ladders to scale the walls, led our soldiers and the main army to climb the walls, and captured the city.)
mbart-large-50	于是,率八十人潜入,盗竖梯登城,擒我前队兵解至。(Then he led eighty men to sneak in, secretly set up ladders to scale the walls, captured our vanguard soldiers, and brought them back.)
m2m100-418M	于是,前队诸臣及八十人即乘夜入城,潜竖梯登城,召我众兵登城克之, (Then the ministers of the vanguard and the eighty men entered the city at night, secretly set up ladders to scale the walls, summoned our troops to climb the walls and capture the city,)
m2m100-1.2B	于是,前行之臣八十人即往,乘夜劫梯,登城,率我众兵登城克之。(Then the leading ministers and the eighty men went at once, seized ladders by night, scaled the walls, led our troops to climb the walls and capture the city.)
nllb-200-600M	由此前出众大臣等,率八十人星夜进城,潜梯登城,率我众兵士驾追夺之, (The prominent ministers who had marched forth led eighty men to enter the city overnight, secretly used ladders to scale the walls, led our soldiers to pursue and capture it,)
nllb-200-1.3B	于是,前行诸臣率八十人速往,乘夜入城,潜登梯登城,率我众兵前进, (Then the leading ministers led eighty men to hasten forward, entered the city at night, secretly climbed ladders to scale the walls, and led our troops forward,)
Q wen	于是，城内诸臣率众出，夜行至近处，围城而攻。我军迎击之，战于城下，败其兵。(Then the ministers in the city led their men out, marched close by night, besieged and attacked the city. Our army met them, fought beneath the walls, and defeated their troops.)
DeepSeek	前行众臣率八十人即往，夜入其城，放火呐喊，登城。遂率我军众兵登城克之。(The leading ministers led eighty men to go at once, entered their city at night, set fire and shouted, then scaled the walls. They then led our troops to climb the walls and capture the city.)

Table 8: Manchu → Classical Chinese translation model predictions. English translations are given in parentheses following the Chinese sentences.

Model	Content
Source	aisi de dosifi dergi de holtorongge, emu beyei bayan elgiyen ojoro jalin, weile muribufi waka ojoro jobolon tumen irgen de isinambi kai,
Target	贪利欺上，止为一身富足，倒置是非，殃及万民矣。(Being greedy for profit and deceiving one's superior, seeking only personal wealth, reversing right and wrong, thus bringing disaster to thousands of people.)
Transformer	入设排称，彼独奉帝德安一万；谁然身S护为民事也。(He entered and arranged claims, alone upholding the emperor's virtue to secure ten thousand; who would protect the people's affairs with his own body S.)
MarianMT	to,,,,
mbart-large-cc25	或贪得而入,只身富躯为图于上,故而启衅也。以致成事,不仅天鉴之, (Or entering out of greed, seeking personal wealth to flatter the superior, thus provoking conflict. As a result, even heaven witnessed it.)
mbart-large-50	贪得利己,蒙蔽其上,以享富足,罪孽及祸患万民。(Being greedy and self-serving, deceiving one's superior to enjoy wealth, committing sins and bringing disaster to thousands of people.)
m2m100-418M	爱西进而上,一身为富足可嘉,罪孽枉费,祸将万民矣。(Loving to advance upward, seeking personal wealth as praiseworthy, committing futile sins and bringing disaster to thousands of people.)
m2m100-1.2B	入贪得之地,君轻敌,实乃为一己之富足,罪孽深重,祸万民也。(Entering a place of greed, the ruler underestimates the enemy, merely seeking personal wealth, committing grave sins and bringing disaster to thousands of people.)
nllb-200-600M	利己欺者,一身富足为利,一事误,致祸祸于万民也。(The selfish and deceptive seek personal profit; one mistake brings disaster to thousands of people.)
nllb-200-1.3B	贪得利己,欺上越,一身富足安逸,是为祸患万民也。(Being greedy and self-serving, deceiving and overstepping one's superior, living in comfort and wealth, thus bringing disaster to thousands of people.)
Qwen	若入边内被欺，虽一身富贵，然遭万民唾弃，非丈夫所为也。(If one is deceived inside the border, even with wealth and rank, one is despised by the people — this is not the act of a man.)
DeepSeek	贪利而欺上，为图一己富贵，酿成罪孽，祸及万民。(Being greedy for profit and deceiving one's superior, seeking personal wealth and rank, committing sins and bringing disaster to thousands of people.)

Table 9: Manchu → Classical Chinese translation model predictions. English translations are given in parentheses following the Chinese sentences.

E More Case for the Conflicts Between Automatic and Human Evaluation

For Case 1 (first line in Table 10), the model's translation completely preserves the original meaning while creatively restructuring the syntax in a manner idiomatic to Manchu (e.g., rendering "视...如..." as "be adali gvnime", and decomposing "遂愤而兴师" into "jili banjifi qouha jihe"). This idiomatic paraphrase resulted in low n-gram overlap with the reference translation, yielding a BLEU-4 score of only 10.24. Yet, human evaluation awarded perfect scores of 5 across all three dimensions—Fidelity, Fluency, and Linguistic Normativity—deeming it a more natural and authentic Manchu expression. This indicates that for highly concise languages like Classical Chinese, excellent translation often requires moving beyond literal alignment—a strategy heavily penalized by match-based metrics like BLEU.

For Case 2 (second line in Table 10), the error lies in a single word: "takahakv" (did not recognize) was mistranslated as "buhekv" (did not give). This changes the sentence's meaning from "I did not recognize the Khan's envoy" to "I did not give (something to) the Khan's envoy," completely reversing the semantics. Despite this serious verb misuse, because the subject, object, and possessive structures of the sentence are entirely correct, the BLEU-4 score remains as high as 75.98. In human evaluation, Fidelity was decisively scored as 2, accurately capturing this fatal error. This exposes BLEU's much higher sensitivity to the main sentence framework (subject-object) than to the accuracy of the predicate verb.

For Case 3 (third line in Table 10), the model's output contains an error in "da hvwa dou" (translated as "大湖岛" instead of "大花岛") and—more critically—completely distorts the meaning of the latter half of the sentence: the original phrase "umai akv waliyafi goidahabi" ("because the inhabitants had long fled, nothing was gained") is rendered as "搜掠已久" ("have been plundering for a long time"). This fundamentally reverses the cause-and-effect relationship and the outcome of the event (from "gaining nothing" to "long-term looting"), constituting a severe distortion of historical facts. However, because the first part of the sentence ("奉圣汗命，往略...与明相邻之地") is translated with high accuracy, and the word "搜掠" is semantically related to "略", the prediction still achieves a BLEU-4 score of 26.03—automatic metrics do not flag it as a complete failure. In contrast, human evaluation decisively assigns low scores of 2 for Fidelity and 2 for Fluency, accurately reflecting the major failure in conveying the core information.

F Background Knowledge

F.1 Manchu Language and Transcription Scheme

As one of the official languages of China's Qing dynasty, Manchu carries a vast corpus of precious historical documents. Manchu belongs to the Tungusic branch of the Altaic language family. Its traditional script evolved from Mongolian and employs a vertical, connected writing style where glyphs vary depending on their position within a syllable. In contemporary academic research and digital preservation contexts, Manchu is often Romanized via transcription for ease of use. This involves using the 26 Latin letters without diacritics to achieve a complete and reversible mapping of Manchu phonemes, ensuring no phonetic information is lost and allowing for restoration. Among various Manchu transcription schemes, the Roman transcription is the most widely used due to its standardized and user-friendly nature, and is uniformly adopted in this paper.

As a supplementary illustration to Table 2, we present the alignment of the vertical Manchu script with Classical Chinese in Figure 9.

F.2 Introduction to Manwen Laodang

Manwen Laodang is a historical text of immense core value from the Qing dynasty and one of the earliest extant Qing documents. Its core content derives from transcriptions of the "Circle-free Character Archives" (i.e., the Old Manchu Archives) formed during the reigns of Emperor Taizu (Nurhaci) and Emperor Taizong (Hong Taiji) before the Qing entered China proper. Its records span from nine years before the Tianming era to the first year of Chongde (1636), covering politics, military affairs, economy, culture, ethnicity, religion, and many other aspects of the Qing founding period. The complete work consists of 180 volumes, divided into three parts: Tianming (81 volumes), Tiancong (61 volumes), and Chongde (38 volumes). The Taizu section corresponds to Tianming, while the Taizong section comprises the 61 volumes of Tiancong and the 38 volumes of Chongde.

Source	Target	Model Prediction
而视我如鸟兽，竟无一言回报，遂愤而兴师。(Regarding us as beasts and birds, they offered no reply at all, and thus, in anger, I raised an army.)	mimbe gasha gurgu -i gese obufi, karu emu gisun jaburakv de korsofi dailambi dere,	mimbe gasha gurgu be adali gvnime karu emu gisun jaburakvngge turgunde, uthai jili banjifi qouha jihe,
我未识为汗之使臣。(I did not recognize him as the Khan's envoy.)	bi han -i elqin be takahakv,	bi han -i elqin be buhekv,
enduringge han -i hesei pi dou, da hvwa dou, yvn cung dou, tiyei xan, ere teisu nikan de aqame tehe ba be tabqilabuqi, umai akv waliyafi goidahabi	奉圣汗命，往略皮岛、大花岛、云从岛、铁山一带与明国相邻之地，因其地居民弃之日久，一无所获 (By the order of the Holy Khan, we marched to plunder the areas adjacent to Ming China around Pidao Island, Dahua Island, Yuncong Island, and Tieshan. However, as the people there had abandoned the land long ago, we found nothing.)	奉圣汗谕旨,皮岛、大湖岛、云从岛、铁山一带,其与明相邻之地方,搜掠已久 (By the edict of the Holy Khan, the regions around Pidao Island, Dahu Island, Yuncong Island, and Tieshan, which are adjacent to Ming, had been plundered for a long time.)

Table 10: Three typical cases of Manchu-Classical Chinese translation. English translations are given in parentheses following the Chinese sentences.

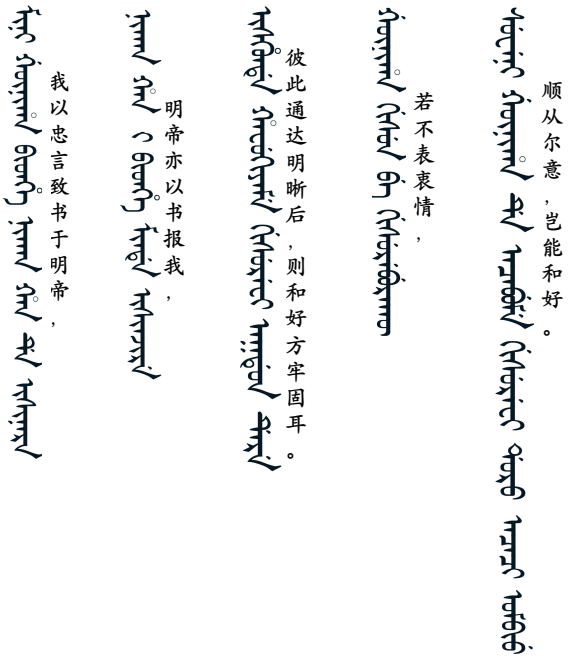

Figure 9: Alignment between Manchu and Classical Chinese sentences.

F.3 Introduction to Classical Chinese

Classical Chinese was the dominant written form of ancient Chinese, modeled on the language of the pre-Qin and Han dynasties, and exhibits significant differences in vocabulary and grammar from modern Chinese. For over two millennia, it served as the primary medium for official documents, scholarly works, and canonical texts. The vast majority of extant ancient Chinese texts are written in Classical Chinese, making it the most important linguistic vessel for historical document information.