Incorporating Lexicon-Aligned Prompting in Large Language Model for Tangut–Chinese Translation

Yuxi Zheng[1] and **Jingsong Yu**[2]
[1]Department of Chinese Language and Literature, Peking University, China
[2]School of Software and Microelectronics, Peking University, China
zhengyuxi@stu.pku.edu.cn
yjs@ss.pku.edu.cn

Abstract

This paper proposes a machine translation approach for Tangut–Chinese using a large language model (LLM) enhanced with lexical knowledge. We fine-tune a Qwen-based LLM using Tangut–Chinese parallel corpora and dictionary definitions. Experimental results demonstrate that incorporating single-character dictionary definitions leads to the best BLEU-4 score of 72.33 for literal translation. Additionally, applying a chain-of-thought prompting strategy significantly boosts free translation performance to 64.20. The model also exhibits strong few-shot learning abilities, with performance improving as the training dataset size increases. Our approach effecttively translates both simple and complex Tangut sentences, offering a robust solution for low-resource language translation and contributing to the digital preservation of Tangut texts.

1 Introduction

The Tangut script, an intricate logographic writing system developed by the Tangut people in the 11th century, served as the official script of the Western Xia dynasty (1038–1227 CE). As a vital cultural artifact, Tangut texts encompass extensive historical, religious, and sociopolitical insights into this once-flourishing Silk Road civilization (Sun, 2023). Despite its scholarly significance, the decipherment and translation of Tangut texts remain formidable challenges. The script's structural complexity, lack of continuous usage traditions, and scarcity of parallel corpora have hindered efficient scholarly access to these invaluable historical records (Kong, 2018). Traditional translation methodologies, reliant on manual "four-line aligned translation" (comprising original text, phonetic transcription, literal translation, and idiomatic translation), demand specialized expertise and labor-intensive efforts, severely limiting the scalability of Tangut studies.

Recent advances in natural language processing (NLP), particularly the emergence of large language models (LLMs), offer unprecedented opportunities to automate low-resource language translation tasks (Lu, 2025). However, existing research has yet to address Tangut translation systematically. Prior work has focused on dictionary compilation, such as A Concise Tangut-Chinese Dictionary (Li, 2012), and manual text analysis, leaving a critical gap in computational methods tailored for Tangut's unique linguistic characteristics. The absence of machine translation systems for Tangut-Chinese conversion underscores both the urgency and innovation potential of this research.

This paper presents the first systematic study on neural machine translation for Tangut texts, targeting two critical tasks: literal translation (character-to-character alignment) and idiomatic translation (semantic restructuring into fluent Chinese). Below are two example sets. For each set: The first line contains the original Tangut text. The second line provides the Chinese character-by-character translation. The third line offers the English character-by-character translation. We performed word-level alignment for the first three lines. The fourth line presents the idiomatic Chinese translation. The fifth line gives the idiomatic English translation.

(1) Tangut

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
愿	永	缘	同	道种	为
wish	forever	pratyaya	same	wisdom-seed	BE

'愿永同缘为道种'
'May we always share the same pratyaya as a seed of the wisdom'

(2) Tangut

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
实	正	法	中	入	当	能
truly	correct	dharma	in	enter	should	be.able

'当能真正法中入'
'we should be able to enter the truly correct dharma'

Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025, pages 228–237
May 4, 2025.

Our work addresses three core challenges: (1) the extreme scarcity of parallel Tangut-Chinese data, (2) the need for precise alignment with authoritative lexicons, and (3) the requirement to adapt Tangut syntax to Classical Chinese expressions. To overcome these barriers, we propose an expert knowledge-enhanced LLM framework that integrates domain-specific dictionaries and chain-of-thought (CoT) prompting strategies. By fine-tuning a pre-trained Classical Chinese LLM (QwenClassical) with carefully curated Tangut datasets, our system achieves robust performance in both translation modes.

Our contributions are threefold:

- **Resource Development**: We compile and release the first publicly available Tangut-Chinese parallel corpus, derived from the *Three Generations Illuminated Collection* and *Avataṃsaka Sūtra*, with 569+525 sentence pairs annotated for literal and idiomatic translation.

- **Methodological Innovation**: We design a hybrid approach combining dictionary-guided character alignment and CoT-based semantic restructuring, enabling accurate translation even with minimal training data.

- **Empirical Validation**: Experiments demonstrate state-of-the-art performance, with BLEU-4 scores of 72.33 (literal) and 64.20 (idiomatic). Ablation studies confirm the effectiveness of domain-adapted LLMs and CoT prompting in low-resource scenarios.

This work not only advances the computational analysis of Tangut texts but also establishes a replicable framework for other under-resourced historical languages. By bridging the gap between ancient script studies and modern NLP, our system empowers historians and linguists to explore Tangut heritage with unprecedented efficiency, fostering new insights into the multicultural dynamics of medieval Eurasia.

The remainder of this paper is organized as follows: Section 2 reviews related work in Tangut linguistics and low-resource machine translation. Section 3 details our methodology, including data preparation and model architecture. Sections 4–5 present experimental results and case analyses, followed by discussions of limitations and future directions.

2 Related Work

This section reviews prior research in two key areas relevant to our study: (1) Tangut linguistics and script decipherment, and (2) low-resource machine translation, with a focus on historical and under-resourced languages. By situating our work within these domains, we highlight the unique challenges and opportunities of applying modern NLP techniques to Tangut texts.

2.1 Tangut Linguistics and Script Decipherment

The Tangut script, also known as Fanwen or Xixia script, is a logographic system comprising over 6,000 characters, developed under the Western Xia dynasty. Early efforts to decipher Tangut texts began in the 20th century, spearheaded by scholars such as Nikolai Nevsky (1960) and Luo Fuchang, who laid the groundwork for understanding its phonetic and semantic structures.

Recent advances in Tangut linguistics have focused on phonology, grammar, and textual analysis. For instance, studies have elucidated the script's phonetic components and syntactic patterns, enabling more accurate transcriptions and translations. The "four-line aligned translation" method, widely adopted in Tangut studies, exemplifies the meticulous process of converting Tangut texts into modern Chinese. This method involves four steps: (1) presenting the original Tangut text, (2) providing a phonetic transcription, (3) generating a literal translation, and (4) producing an idiomatic translation. While effective, this approach is labor-intensive and heavily reliant on expert knowledge, underscoring the need for computational solutions.

Despite these advancements, the field faces persistent challenges, including the scarcity of parallel corpora, the ambiguity of Tangut characters, and the lack of standardized tools for automated analysis (Liu, 2022). These limitations have hindered the scalability of Tangut research, making it an ideal candidate for NLP-driven innovations.

2.2 Low-Resource Machine Translation

Machine translation (MT) for low-resource languages has gained significant attention in recent years, driven by the success of neural models and transfer learning techniques. Early approaches relied on rule-based and statistical methods, which struggled to handle the morphological and syntactic complexities of under-resourced languages. The advent of neural machine translation, particularly sequence-to-sequence models and transformer architectures, has revolutionized the field, enabling more robust and context-aware translations (Zoph, 2016).

For historical and ancient languages, MT systems must address unique challenges, such as incomplete lexicons, fragmented texts, and the absence of native speakers. Recent work has demonstrated the potential of LLMs in this domain (Jiao et al, 2023). For example, BERT-based models have been adapted for Classical Chinese(Yu et al, 2020), while GPT variants have been fine-tuned for ancient Greek (Lu et al, 2025) and Latin (Stüssi et al, 2024) . These models leverage pre-training on large corpora and domain-specific fine-tuning to achieve state-of-the-art performance.

A key innovation in low-resource MT is the use of auxiliary resources, such as dictionaries, parallel texts, and multilingual embeddings (Ammar et al, 2016), to enhance model performance. Techniques like back-translation, data augmentation, and transfer learning (Zoph et al, 2016) have proven effective in scenarios with limited parallel data. Additionally, prompting strategies, including chain-of-thought (CoT) (Wei et al, 2022) and few-shot learning (Wang et al, 2020), have emerged as powerful tools for guiding LLMs in low-resource settings.

Despite these advances, the application of MT to Tangut texts remains unexplored. The script's logographic nature, combined with its historical and cultural specificity, presents unique challenges that require tailored solutions. Our work bridges this gap by integrating domain-specific lexicons and CoT prompting into a fine-tuned LLM framework, enabling accurate and scalable Tangut-Chinese translation.

2.3 Bridging the Gap

By synthesizing insights from Tangut linguistics and low-resource MT, our research addresses a critical gap in both fields. We build on the foundational work of Tangut scholars while leveraging cutting-edge NLP techniques to automate and enhance the translation process. This interdisciplinary approach not only advances Tangut studies but also contributes to the broader field of historical language processing, offering a replicable framework for other under-resourced scripts.

In the following sections, we detail our methodology, which combines expert knowledge with neural models to achieve robust and interpretable translations of Tangut texts.

3 Methodology

Our methodology addresses the dual challenges of translating Tangut texts into Chinese through two interconnected tasks: literal translation (character-level alignment) and idiomatic translation (semantic restructuring). We propose a hybrid framework that integrates domain-specific lexicons with a fine-tuned large language model (LLM), enhanced by chain-of-thought (CoT) prompting strategies. This section details our data preparation, model architecture, and training protocols.

3.1 Data Preparation

3.1.1 Lexical Resources

We utilize the *Concise Tangut-Chinese Dictionary* (Li, 2012), which provides 6,703 Tangut character entries with 8,245 annotated Chinese definitions. Each processed entry includes:

- **Full Definitions** (Dict)**:** Multi-sense explanations with part-of-speech tags (e.g., "1. 種、苗、裔[名詞] 2. 胤 3. 明").
- **Simplified Definitions** (DictSingle): Single-sense keywords derived from Dict (e.g., "种、苗、裔、胤、明").

These definitions serve as lexical anchors for character-level alignment during translation.

3.1.2 Parallel Corpus Construction

We compile a Tangut-Chinese parallel corpus from two primary sources:

- **Three Generations Illuminated Collection**: 569 sentence pairs with four-line aligned translations (original Tangut, phonetic transcription, literal Chinese, idiomatic Chinese), sourced from Sun (2022).
- **Avataṃsaka Sūtra Vol. 77**: 525 sentence pairs, supplemented by ChatGPT-generated literal translations from existing Japanese paraphrases (Arakawa, 2011).

The corpus is split into:

- **Training Set:** 95% of the *Three Generations* data (541 pairs).
- **Test Set:** 5% of the *Three Generations* data (28 pairs).
- **Generalization Test**: the Avataṃsaka Sūtra Vol. 77 data(525 pairs).

3.2 Model Architecture

3.2.1 Base LLM: QwenClassical

We employ QwenClassical, a variant of Qwen1.5-14B-Chat (Bai et al., 2023), pre-trained on 36GB of Classical Chinese texts (e.g., Shiji, Zizhi Tongjian) and fine-tuned with 390K task-specific examples (e.g., classical-modern Chinese translation, poetry generation). The specific training process refers to the existing classical Chinese model (Zhang, 2024). This domain adaptation enables robust handling of Tangut-to-Chinese syntactic and semantic divergences.

3.2.2 Expert Knowledge Integration

To inject Tangut-specific linguistic knowledge, we change each Tangut character with its Dict or DictSingle definitions during input encoding. For the term '[illegible]', when using DictSingle, its prompt is shown in Table 1.

Tangut Script	Prompt
[illegible]	The candidate words for each Tangut Character are The first character: [罪、过] The second character: [非、否、不] The third character: [皆、咸、俱、普、悉、总、极、周、竞] The fourth character: [不] The fifth character: [做、作、为]

Table 1: DictSingle Definitions for the Tangut Characters '[illegible]'

This approach grounds the model in authoritative lexical semantics while preserving contextual flexibility. By explicitly associating each Tangut character with its possible meanings, the model can better disambiguate polysemous characters and generate more accurate translations. Additionally, the use of simplified definitions (DictSingle) reduces noise and improves computational efficiency, as the model focuses on the most relevant semantic information.

3.3 Training Strategy

3.3.1 Literal Translation

We constructed input-output pairs as shown in Table 2 for fine-tuning the model for literal translation.

3.3.2 Idiomatic Translation

Idiomatic translation is framed as a two-step CoT task:

1. **Literal Drafting**: Generate a provisional literal translation $L = \{l_1, \ldots l_n\}$.

2. **Semantic Refinement**: Restructure L into fluent Classical Chinese $Y = \{y_1, \ldots y_n\}$ using in-context examples.

We constructed input-output pairs as shown in Table 3 for fine-tuning the model for Idiomatic translation.

This CoT strategy mimics human translation workflows, reducing semantic drift in low-resource scenarios.

Input	Output
Provide the literal translation of the Tangut script. The candidate words for each Tangut Character are The first character: […] The second character: […] The third character: […] The fourth character: […] The fifth character: […]	The literal translation is: …

Table 2: Input-Output Pairs for Fine-Tuning the Model for Literal Translation

Input	Output
First, provide the literal translation of the Tangut script, and then give the idiomatic translation based on the literal translation. The candidate words for each Tangut Character are The first character: […] The second character: […] The third character: […] The fourth character: […] The fifth character: […]	The literal translation is: … The idiomatic translation is: …

Table 3: Input-Output Pairs for Fine-Tuning the Model for Idiomatic Translation

3.4 Implementation Details

- **Hardware**: 2×NVIDIA A800 GPUs (80GB VRAM).
- **Optimization**: AdamW (learning rate 3e−4, cosine decay), mixed-precision (bfloat16).
- **Training**: 5 epochs, batch size 8, gradient accumulation steps 1.
- **Tokenization**: SentencePiece (32K vocabulary) with Tangut Unicode block extensions.
- **Finetuning**: LoRA finetuning with Zero2 technique.

4 Experiments

This section evaluates the performance of our Tangut-Chinese machine translation system through quantitative metrics, ablation studies, and qualitative analyses. We assess both literal and idiomatic translation tasks, investigate the impact of training data scale, and validate the model's generalization capability across diverse Tangut texts.

4.1 Experimental Setup

4.1.1 Datasets

- **Primary Dataset**: 569 sentence pairs from Three Generations Illuminated Collection, split into 541 training and 28 test pairs.
- **Generalization Dataset**: 525 sentence pairs from Avataṃsaka Sūtra Vol. 77, with 200 held-out pairs for cross-domain evaluation.
- **Lexical Resources**: 6,703 Tangut characters annotated with 8,245 Chinese definitions from A Concise Tangut-Chinese Dictionary.

4.1.2 Baselines and Variants

We compare two base models:

- **Qwen:** Original Qwen1.5-14B-Chat.
- **QwenClassical:** Our pre-trained Classical Chinese variant.

For each model, we test four configurations:

- **Dict**: Full dictionary definitions.
- **DictSingle**: Simplified single-keyword definitions.
- **Prompt-0-shot**: Direct translation instruct-tion.
- **PromptCoT**: Chain-of-thought prompting.

4.1.3 Evaluation Metrics

- **BLEU-4**: Measures n-gram overlap between machine and reference translations (Papineni, 2002).
- **Human Evaluation**: Three Tangut linguistics experts rate translations on a 5-point Likert scale (1: Incoherent, 5: Fluent and Faithful).

4.2 Main Results

4.2.1 Literal Translation Performance

Table 4 compares BLEU-4 scores across configurations. QwenClassical with DictSingle achieves the highest score (72.33), outperforming the base Qwen model by 2.83 points. Simplified definitions (DictSingle) consistently improve performance over full definitions (Dict), likely due to reduced lexical ambiguity.

Model	Dict	DictSingle
Qwen	69.78	71.50
QwenClassical	70.86	**72.33**

Table 4: Performance of Tangut-Chinese Literal Translation on Test Set (BLEU-4)

4.2.2 Idiomatic Translation Performance

Table 5 demonstrates the superiority of CoT prompting (PromptCoT) over direct prompting (Prompt-0-shot), with a 12.14 BLEU-4 improvement. QwenClassical+DictSingle+PromptCoT achieves the best performance (64.20), validating the effectiveness of stepwise semantic restructuring.

Model	Prompt-0-shot	PromptCoT
QwenClassical+Dict	51.06	62.54
QwenClassical+DictSingle	52.58	**64.20**

Table 5: Performance of Tangut-Chinese Idiomatic Translation on Test Set (BLEU-4)

4.3 Impact of Training Data Scale

To assess data efficiency, we vary the training set size from 100 to 500 pairs (Table 6). Both tasks exhibit steady performance growth, with literal translation saturating at ~500 samples (BLEU-4: 73.41). Notably, the model achieves 62.83 BLEU-4 for literal translation with only 100 samples, demonstrating strong few-shot learning capabilities.

Training Set Size	Literal Translation	Idiomatic Translation
100	62.83	59.53
200	70.06	62.34
300	69.57	62.73
400	71.31	65.94
500	**73.41**	**66.05**

Table 6: BLEU-4 Scores with Varying Training Data Sizes

4.4 Cross-Domain Generalization

We evaluate generalization by fine-tuning on incremental subsets of Avataṃsaka Sūtra data (Table 7). With 200 supplementary pairs, the model achieves 30.62 (literal) and 37.00 (idiomatic) BLEU-4 on the out-of-domain test set, confirming its adaptability to new Tangut genres.

Added Pairs	Literal Translation	Idiomatic Translation
40	23.88	30.92
80	24.58	32.62
120	25.45	34.76
160	27.28	35.49
200	**30.62**	**37.00**

Table 7: Generalization Performance on Avatamsaka Sūtra Dataset (BLEU-4)

4.5 Comparison with other high performance models

To clarify the need for fine-tuning, experiments were conducted on ChatGPT-4o, Gemini-2.0-Flash and DeepSeek V3, which currently have excellent comprehensive performance, using a few-shot method. Dictsingle and Dictsingle+PromptCoT is used for the translation. Five samples were randomly selected from the training set as examples and input into the model, and then the BLEU-4 score was calculated on the generated results. The experimental results are summarized in Table 8.

Model	Machine Literal Translation	Machine Idiomatic Translation
ChatGPT-4o	20.13	14.96
DeepSeek V3	38.85	24.33
Gemini-2.0-Flash	32.07	19.68
ours	**72.33**	**64.20**

Table 8: Comparison with other high-performance models (using few-shot methods) (BLEU-4)

From the experimental results, whether it is automatic literal translation or automatic idiomatic translation, the model proposed in this paper scores significantly higher than ChatGPT-4.0, DeepSeek V3, and Gemini-2.0-Flash under few-shot learning methods. This indicates that general models struggle to meet the demands of literal and free translation tasks for Tangut texts due to a lack of relevant content in their training data aimed at the design tasks of this study. However, through fine-tuning, we have significantly improved the model's adaptability to specific tasks, resulting in a substantial increase in the quality of both automatic literal and idiomatic translations. Based on the above comparative results, we can further validate the effectiveness and necessity of fine-tuning strategies.

4.6 Human Evaluation

Three experts rated 50 randomly sampled translations (Table 9). QwenClassical+DictSingle+PromptCoT received the highest fluency (4.12/5) and faithfulness (4.35/5) scores, aligning with automated metrics.

Model	Fluency	Faithfulness
Qwen+Dict	3.45	3.78
QwenClassical +DictSingle +PromptCoT	**4.12**	**4.35**

Table 9: Expert Ratings of Translation Quality (5-point Likert Scale)

4.7 Case Study

To visually demonstrate the effects of automatic translation and automatic interpretation, typical examples are selected for analysis separately. The results are shown in Table 10 and Table 11.

Tangut Script	Reference Literal Translation	Machine Literal Translation
[illegible]	香花布列	香花排列
[illegible]	凡君子者，他利故己不忘，不学者无	夫子者，他利为己不忘，不学者，则无

Table 10: Example of Literal Translation

Tangut Script	Reference Literal Translation	Reference Idiomatic Translation	Machine Idiomatic Translation
[illegible]	此复退难自何见	此复难退自何见	此复遣返难自见
[illegible]	△想△则人悲痛	我等每思则悲哭	我等每思则悲哭

Table 11: Example of Idiomatic Translation

The machine literal translation examples of simple sentences and complex sentences are shown in Table 10. The analysis results show that for the translation of simple sentences, the model can accurately capture the semantic information of the source language and achieve accurate conversion. For the translation of complex sentences, although there are slight differences in local expression between machine translation output and reference translations. Overall, they still maintain a high level of semantic integrity and expression accuracy. This indicates that the model proposed in this study has good robustness in handling translation tasks of different language complexities.

Table 11 presents examples of machine idiomatic translation, where omitted content in the standard translation is represented by the symbol "△". When automatically paraphrased, the model is able to effecttively identify and supplement this implicit information, thus generating a more complete translation.

4.8 Error Analysis

Common failure modes include:

1. **Ambiguous Characters**: Misinterpreting Tangut homographs.

2. **Syntactic Divergence**: Over-literal restructuring (e.g., retaining Tangut SOV order in Chinese SVO contexts).

3. **Cultural References**: Missing context-specific terms (e.g., Buddhist technical vocabulary).

5 Conclusions

This paper presents the first systematic study on neural machine translation for Tangut texts, addressing the critical challenges of translating a historical logographic script with extremely limited parallel resources. By integrating domain-specific lexicons, chain-of-thought prompting, and a pre-trained Classical Chinese LLM, we develop a hybrid framework that achieves robust performance in both literal and idiomatic translation tasks. Our key findings and contributions are summarized as follows:

- Effective Resource Utilization: The integration of expert-curated dictionaries (A Concise Tangut-Chinese Dictionary) with neural models significantly improves translation accuracy, achieving state-of-the-art BLEU-4 scores of 72.33 (literal) and 64.20 (idiomatic). This demonstrates the viability of leveraging domain knowledge to compensate for data scarcity in historical language processing.

- Methodological Innovation: Our two-stage CoT prompting strategy, which decouples literal alignment from semantic restruckturing, mimics human translation workflows and reduces error propagation. Ablation studies confirm that this approach outperforms direct translation by 12.14 BLEU-4 points in idiomatic tasks.

- Practical Impact: The release of the first publicly available Tangut-Chinese parallel corpus (1,094 sentence pairs) and the trained models provides foundational resources for accelerating Tangut studies. Case analyses show that our system can handle complex syntactic divergences and culturally specific references, such as Buddhist terminology in Avataṃsaka Sūtra.

- Broader Implications: This work establishes a replicable framework for other under-resourced historical languages, demonstrating how LLMs can bridge the gap between computational linguistics and philology. The model's strong few-shot learning capability (62.83 BLEU-4 with 100 samples) highlights the potential for scaling to other extinct or low-resource scripts.

6 Limitations and Future Work

While our system marks a significant advance, three limitations warrant further investigation:

- Data Scarcity: Expanding the parallel corpus to include more genres (e.g., legal documents, poetry) and dialects could enhance generalization.

- Context Handling: Current models struggle with long-range dependencies in multi-sentence Tangut texts. Future work should explore document-level translation and multimodal approaches (e.g., integrating textual and glyph features).

- Human-in-the-Loop Optimization: Developing interactive tools for expert validation and error correction would improve practical utility.

By addressing these challenges, we aim to transform Tangut studies from a niche philological endeavor into a data-driven interdisciplinary field. Our work underscores the transformative potential of NLP in preserving linguistic heritage and fostering cross-cultural understanding.

References

Bojun Sun. 2023. The Current Situation and Future of Literature Research in the Western Xia Regime. *Journal of Southwest Minzu University (Humanities and Social Science)*, 44(1):14-21.

Xianghui Kong. 2018. The Construction and Research of Tangut Corpus from the Perspective of Corpus. *Northwestern Journal of Ethnology*, (4):199-205.

Arakawa Shintaro. 2011. Annotated Translation of the Princeton University Collection of Tangut Huayan Sutra Chapter 77. *Journal of Asian and African Studies*, 81:147-305.

Changqing Liu. 2022. *Digital Research on Tangut Script*. Guangzhou: Sun Yat-sen University Press.

Bojun Sun. 2022. *Compilation and Study of Yuan Dynasty Baiyun Sect Tangut Documents*. Beijing: China Social Sciences Press.

Fanwen Li. 2012. *Concise Tangut-Chinese Dictionary*. Beijing: China Social Sciences Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Weijing 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.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. *arXiv preprint arXiv*:2301.08745v4.

Jinze Bai, Shuai Bai, Yunfei Chu, et al. 2023. Qwen Technical Report. *arXiv preprint arXiv*:2309.16609.

Yuyan Zhang. 2024. *Research and Design of Ancient Chinese Large Language Models*. Master's thesis, Peking University.

Kaiwen Lu, Yating Yang, Fengyi Yang, Rui Dong, Bo Ma, Aihetamujiang Aihemaiti, Abibilla Atawulla, Lei Wang, and Xi Zhou. 2025. Low-Resource Language Expansion and Translation Capacity Enhancement for LLM: A Study on the Uyghur. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8360–8373, Abu Dhabi, UAE. Association for Computational Linguistics.

Nevsky Nikolai. 1960. *Tangut Philology: Research and Dictionary*. 2 vols. Moscow: Oriental Literature Publishing House.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Yu Peng and Wang Xin. 2020. BERT-Based Named Entity Recognition in Chinese Twenty-Four Histories. In *Proceedings of the International Conference on Web Information Systems and Applications*. Cham: Springer International Publishing.

Elina Stüssi and Phillip Ströbel. 2024. Part-of-Speech Tagging of 16th-Century Latin with GPT. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 196–206, St. Julians, Malta. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Yaqing Wang, Quanming Yao, James Kwok, and Lionel M. Ni. 2020. Generalizing From a Few Examples: A Survey on Few-Shot Learning. *ACM Computing Surveys*, 53(3):1–34.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv*:1602.01925.

A Examples of Translation Results

Examples of Tangut-to-Chinese Literal Translation (Closed Test)

Tangut Script:
[illegible]
Reference Literal Translation:
盛衰孰知何长短，名利实虚不殆谓
Machine Literal Translation:
盛忍谁知何长短，名利实虚不怠谓。

Tangut Script:
[illegible]
Reference Literal Translation:
菩提道上山关无，般若眼内翳何熏。
Machine Literal Translation:
报日道上山关无，瞿昙眼内翳何熏。

Tangut Script:
[illegible]
Reference Literal Translation:
家宅宝钱九亲及，禄福于至弃毁者。
Machine Literal Translation:
家宅宝财九亲及，禄福上至弃毁者。

Tangut Script:
[illegible]
Reference Literal Translation:
性喜旱因多乃说，△降皆惧毁。
Machine Literal Translation:
性爱旱因多乃说，△雨皆消毁。

Tangut Script:
[illegible]
Reference Literal Translation:
若复实不说，一时名利思。
Machine Literal Translation:
若复实不说，一时名利想。

Examples of Tangut-to-Chinese Literal Translation (Open Test)
Tangut Script:
[illegible]
Reference Literal Translation:
凡君子者，他利故已不忘，不学者无
Machine Literal Translation:
夫子者，他利为己不忘，不学者，则无

Tangut Script:
[illegible]
Reference Literal Translation:
己利故他不绝，不教亦无。
Machine Literal Translation:
自利依他不舍，不教亦无。

Tangut Script:
[illegible]
Reference Literal Translation:
金真铃铎云如布
Machine Literal Translation:
金真铃铎云如偈

Tangut Script:
[illegible]
Reference Literal Translation:
宝枝杂布严好为
Machine Literal Translation:
宝支杂行长卷为

Tangut Script:
[illegible]
Reference Literal Translation:
宝王间错严饰为
Machine Literal Translation:
宝王混行金刚为

Examples of Tangut-to-Chinese Idiomatic Translation (Closed Test)
Tangut Script:
[illegible]
Reference Idiomatic Translation:
盛衰孰知何短长？名利实虚不懈怠。
Machine Idiomatic Translation:
盛忍孰知何长短？名利实虚不怠谓。

Tangut Script:
[illegible]
Reference Idiomatic Translation:
菩提道上无山险，般若眼内无翳熏。
Machine Idiomatic Translation:
报日道上无关山，瞿昙眼内翳熏何。

Tangut Script:
[illegible]
Reference Idiomatic Translation:
家宅宝财和九亲，上至福禄均毁弃。
Machine Idiomatic Translation:
家宅宝财及九亲，福禄至弃毁家宅。

Tangut Script:
[illegible]
Reference Idiomatic Translation:
因大旱多言性喜，降雨皆惧毁。
Machine Idiomatic Translation:
性爱旱依多乃说，而雨皆毁坏。

Tangut Script:
[illegible]
Reference Idiomatic Translation:
若复不实说，一时名利思。
Machine Idiomatic Translation:
若复不说实，一时名利想。

Examples of Tangut-to-Chinese Idiomatic Translation (Open Test)
Tangut Script:
[illegible]
Reference Idiomatic Translation:
凡君子者，利他故不忘己，无不学者；
Machine Idiomatic Translation:
凡君子者，于他利故，己忘不忘，不学者无；

Tangut Script:
[illegible]
Reference Idiomatic Translation:
利己故不绝他，亦无不教。
Machine Idiomatic Translation:
为利自己故不断他，亦不施教。

Tangut Script:
[illegible]
Reference Idiomatic Translation:
真金铃铎如云布
Machine Idiomatic Translation:
真金铃铎如云布

Tangut Script:
[illegible]
Reference Idiomatic Translation:
宝枝杂布为严好

Machine Idiomatic Translation:
宝枝杂布好严密

Tangut Script:
[illegible]
Reference Idiomatic Translation:
宝王间错为严饰
Machine Idiomatic Translation:
宝王间错严饰为