BrailleLLM: Braille Instruction Tuning with Large Language Models for Braille Domain Tasks

Tianyuan Huang[1*], **Zepeng Zhu**[1*], **Hangdi Xing**[2*†], **Zirui Shao**[1]
Zhi Yu[1,4†], **Chaoxiong Yang**[1], **Jiaxian He**[1], **Xiaozhong Liu**[3], **Jiajun Bu**[1]
[1] Zhejiang Key Laboratory of Accessible Perception and Intelligent Systems, Zhejiang University
[2] Alibaba Group [3] Worcester Polytechnic Institute
[4] Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and DataSecurity
{huangty, xinghd, yuzhirenzhe}@zju.edu.cn

Abstract

Braille plays a vital role in education and information accessibility for visually impaired individuals. However, Braille information processing faces challenges such as data scarcity and ambiguities in mixed-text contexts. We construct English and Chinese Braille Mixed Datasets (EBMD/CBMD) with mathematical formulas to support diverse Braille domain research, and propose a syntax tree-based augmentation method tailored for Braille data. To address the underperformance of traditional fine-tuning methods in Braille-related tasks, we investigate Braille Knowledge-Based Fine-Tuning (BKFT), which reduces the learning difficulty of Braille contextual features. BrailleLLM employs BKFT via instruction tuning to achieve unified Braille translation, formula-to-Braille conversion, and mixed-text translation. Experiments demonstrate that BKFT achieves significant performance improvements over conventional fine-tuning in Braille translation scenarios. Our open-sourced datasets and methodologies establish a foundation for low-resource multilingual Braille research[1].

1 Introduction

Braille is a tactile writing system essential for the visually impaired, providing a crucial means of accessing textual information. While speech synthesis has improved accessibility, Braille remains vital in areas like education and leisure reading for the blind (Johan Rempel and CATIS, 2022; Awang et al., 2024; Kana and Hagos, 2024). However, the scarcity of Braille resources and challenges in Braille comprehension have adversely impacted the well-being of the visually impaired community, limiting their access to education. Braille text translation technology offers a promising solution, as illustrated in Fig. 1. Existing Braille research typically focuses on isolated tasks (Wang et al., 2019; Huang et al., 2023; Yu et al., 2023), while visually impaired individuals often need to process documents containing mixed content, including text, formulas, and symbols. Furthermore, the scarcity and limited accessibility of Braille resources further hinder progress in this field.

The primary challenge in mixed-content Braille tasks lies in the inherent ambiguity of Braille-to-text conversion. Braille consists of only 64 distinct Braille cells, typically represented by corresponding Braille ASCII codes in computing systems, which map to multiple plain-text counterparts, including Chinese characters, English words, punctuation, and mathematical symbols. This inherent ambiguity enables a single Braille segment to carry diverse semantic interpretations. For instance, a Chinese Braille segment may correspond to over a dozen Chinese characters. Such a polysemous nature necessitates learning contextual features from extensive training data for robust Braille-Chinese translation.

Braille translation research has progressed from rule-based and statistical methods to neural machine translation (NMT) (Huang et al., 2023; Yu et al., 2023), substantially improving the learning capacity for Braille sequence patterns. Nevertheless, studies remain inadequate in addressing the mutual conversion of hybrid-content documents due to Braille translation's inherent complexity and resource limitations. In low-resource scenarios, transferring generic semantic knowledge from large-scale models to domain-specific tasks has demonstrated effectiveness (Chowdhery et al., 2023; Touvron et al., 2023).

This paper presents BrailleLLM, a framework that employs instruction-tuned large language models to resolve Braille domain challenges. We construct a comprehensive Braille corpus and design an instruction fine-tuning strategy to en-

[*] Equal contribution.
[†] Corresponding author.
[1] https://github.com/Tianyuan-Huang/BrailleLLM

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28590–28601
November 4-9, 2025

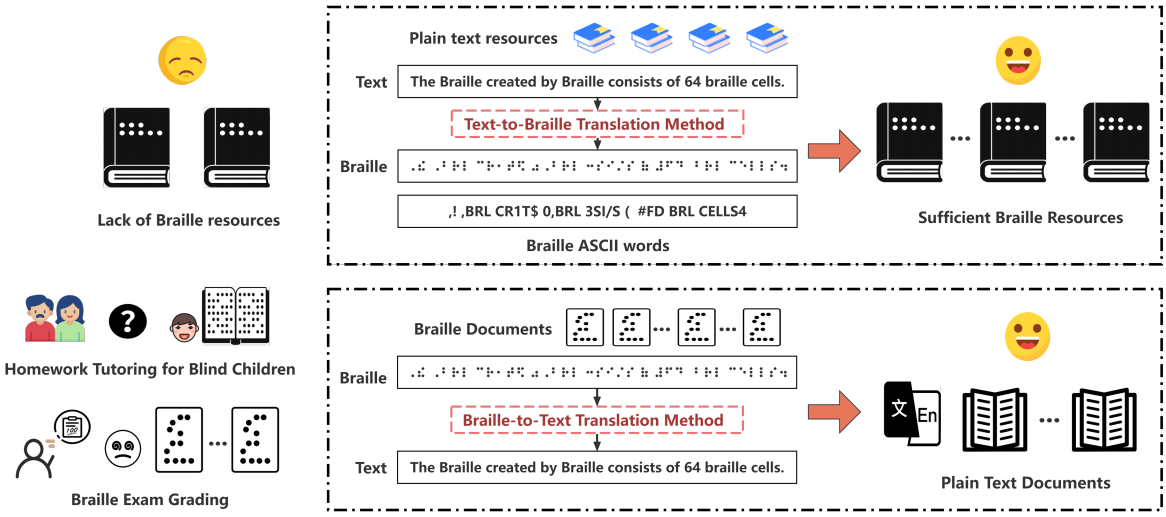

Figure 1: Braille translation methodologies enable the conversion of plain-text resources into Braille or Braille ASCII words (the computational representation of Braille), granting visually impaired populations access to Braille materials. Furthermore, Braille-to-text translation technology enhances educational accessibility by facilitating home tutoring for blind children and empowering Braille school educators to evaluate complex Braille examination papers efficiently.

hance model performance. Specifically, we curate task-specific instruction templates and create hybrid Braille-text datasets, including the Chinese Braille Mixed-text Dataset (CBMD) and English Braille Mixed-text Dataset (EBMD). While powerful, LLMs treat Braille as opaque token sequences, forcing them to inefficiently learn its fundamental syntactic rules from scarce data during fine-tuning. We introduce Braille Knowledge-Based Fine-Tuning (BKFT) to overcome this by directly injecting these structural priors. BKFT thus redirects the model's capacity from rediscovering low-level rules to mastering high-level semantic disambiguation and translation.

Our main contributions can be summarized as follows:

- We developed the CBMD and EBMD datasets for mixed-text Braille conversion, along with corresponding instruction templates. These resources comprehensively cover various Braille-related tasks.
- Based on the unique characteristics of Braille data, we propose a data augmentation method utilizing constituent and dependency syntax trees. This approach effectively increases dataset diversity while maintaining data validity.
- We introduce a fine-tuning method based on a Braille prior knowledge base, enabling large general-purpose models to acquire Braille-specific prior knowledge. This method provides a universal foundational approach for encoding Braille data.

2 Related Work

2.1 Braille Information Processing

Braille information processing research focuses on bidirectional conversion between natural languages and Braille systems. Conventional approaches primarily employ rule-based and statistical models (Jiang et al., 2002; Zhang et al., 2022; Minghu et al., 2000; Wang et al., 2017, 2016; Jariwala and Patel, 2017): pinyin knowledge bases (Jiang et al., 2002) and phonographic mapping tables (Wang et al., 2017) have driven the development of Chinese-Braille conversion systems, while mathematical formula transformation relies on predefined symbolic rules (Jariwala and Patel, 2017; Zatserkovnyi et al., 2019; Egli, 2009). With the evolution of NMT (Cai et al., 2019; Jiang et al., 2021; Wang et al., 2019; Huang et al., 2023; Yu et al., 2023), researchers have developed end-to-end Braille conversion models. Although NMT demonstrates adaptability in multi-lingual scenarios (Vaswani, 2017; Zhou et al., 2022; Hu et al., 2021; Lu et al., 2022), its dependence on large-scale annotated data inherently conflicts with the scarcity of Braille corpora. Current studies predominantly concentrate on single-content conversion, yet fail to address the critical

challenge of mixed-content processing faced by visually impaired communities, resulting in limited applicability of existing methods in real-world scenarios.

2.2 Low resource translation

Research on low-resource translation primarily addresses data scarcity and encoding complexity through transfer learning, meta-learning, and data augmentation. Transfer learning enhances model adaptability via pre-training and fine-tuning paradigms (Di Gangi and Federico, 2017; Qi et al., 2018; Dong, 2023; Gao et al., 2024). Meta-learning approaches optimize parameter initialization and generalization through cross-domain knowledge sharing (Li et al., 2020; Park et al., 2020). Data augmentation techniques synthesize parallel corpora via back-translation (Sennrich, 2015; Bawden et al., 2019; Sánchez-Martínez et al., 2020), syntax-rule generation (Lucas et al., 2024), and bilingual templates (Li et al., 2024), with syntax-tree strategies offering novel paradigms for morphologically complex languages (Lucas et al., 2024). However, existing Braille translation studies have not effectively leveraged knowledge transfer methodologies, and current data augmentation approaches lack tailored solutions for Braille-specific data characteristics.

2.3 Large Language Models

LLMs have revolutionized natural language processing through the self-supervised pretraining paradigm based on Transformer architectures. Pioneering works like BERT (Devlin, 2018) and GPT series (Radford, 2018; Radford et al., 2019; Brown et al., 2020) models demonstrated the efficacy of the pretrain-finetune paradigm in cross-task transfer. Continued model scaling and architectural innovations have enhanced LLMs' generalization capabilities (Chowdhery et al., 2023; Touvron et al., 2023). Despite contemporary LLMs demonstrate general task-solving capabilities (Liu et al., 2024; Achiam et al., 2023), their applications in specialized domains remain constrained by domain-specific knowledge gaps. Existing domain-adapted solutions (e.g., Med-PaLM (Singhal et al., 2023) and LawLLM (Shu et al., 2024)) enhance task performance through instruction tuning and knowledge integration. Nevertheless, current general-purpose LLMs exhibit inadequate Braille processing capabilities and lack Braille-specific adaptations.

3 Dataset

We present the first comprehensive multi-lingual Braille dataset suite for Chinese and English Braille processing tasks, addressing the critical data scarcity challenge in Braille NLP. Our dataset construction integrates three key components: (1) a Chinese Braille Dataset (CBD), (2) a Chinese Braille Mixed-text Dataset (CBMD), and (3) an English Braille Mixed-text Dataset (EBMD). To enable instruction tuning for large language models (LLMs), we generate task-specific datasets through diverse instruction templates. Given the high annotation cost of Chinese Braille, we propose a novel syntax-aware data augmentation method tailored to its linguistic properties.

3.1 Data Collection

The collection of Braille datasets has faced significant challenges, especially in acquiring parallel datasets between Plain text and Braille texts. Blind schools and Braille libraries often use Braille, but structured parallel data is difficult to obtain and verify. As a result, we first collected plain texts, then used conversion tools to generate an initial Braille version, followed by manual correction and annotation. For the CBD, we selected three months of standardized Chinese news articles from the People's Daily. The CBMD construction focuses on STEM education needs, utilizing junior and senior high school mathematics documents containing Chinese text and mathematical formulas. For EBMD, we adapt the open-source NuminaMath-CoT [2] dataset, comprising English mathematical problems from online exams and forums. Raw texts undergo rigorous preprocessing, including deduplication, format normalization using LaTeX standardization for mathematical formulas, sanitization to filter non-standard characters, and error correction through rule-based verification. Dataset statistics and sample structures are detailed in Table 1.

3.2 Data Annotation

The acquisition of initial Braille was obtained by utilizing text-to-Braille and formula-to-Braille conversion libraries through concatenation to generate hybrid Braille. For Chinese mixed text-to-Braille conversion, dedicated APIs corresponding to Chinese and mathematical formulas from the China

[2]https://huggingface.co/datasets/AI-MO/NuminaMath-CoT

Data type	Data volume	Data Example
Pure Chinese	11000	该增长水平使手机市场成为当今最活跃的商品市场之一。
Mixed Chinese	11000	答：图1中阴影部分面积为：$\frac{3}{m+1}$ 平方米
Mixed English	11000	Suppose that $g(x)=5x-3$. What is $g^{-1}(g^{-1}(14))$?

Table 1: Data sample and statistics.

Braille Digital Platform [3] were employed. English hybrid Braille generation utilized the English and mathematical formula conversion interfaces in the liblouis library [4].

Despite the automatic conversion, errors and inconsistencies remained, including incorrect word segmentation, tonal marking errors, and punctuation issues in Chinese Braille, as well as translation errors in both English words and formulas. We identified various errors and compiled detailed annotation guidelines for both Chinese and English Braille. A team of 10 Braille domain experts and AI specialists was tasked with the annotation and quality control process. Multiple rounds of validation were performed to ensure the accuracy of the final datasets. The resulting datasets, CBD, CBMD, and EBMD, provide word-level alignment between plain text and Braille text.

To support additional Braille domain tasks, such as Braille word segmentation and conversion between Braille and Pinyin, we expanded the datasets to include specific task data, as illustrated in Table 2. Furthermore, we developed dozens of instruction templates and corresponding instruction datasets, as shown in Figure 2.

3.3 Braille Data Augmentation

The creation of Chinese Braille datasets is notably more complex and incurs significantly higher verification costs compared to English Braille, due to the highly ambiguous character representations in Chinese Braille, whereas English words and Braille exhibit clear correspondences. To address this challenge, we investigate data augmentation methods tailored to the semantic composition characteristics of Chinese Braille texts.

By performing constituency syntax tree parsing on Braille texts, we categorize Braille segments into dozens of attribute types, including quantifier attributes, pronoun attributes, and location attributes. For example, the Braille sequence "IW N%K*AD K5AH)1G$A T1QUAL5 Q72H<AD LI'L3" corresponds to the constituency syntax tree shown in Figure 3. The NP (noun phrase) component comprises a quantifier Braille segment "IW", an adjective Braille segment "N%K*AD", and a noun Braille segment "K5AH)1G$A". The interdependencies among Braille segments of different attributes within the syntax tree ensure the linguistic validity of Braille texts.

We construct a knowledge base encompassing various Braille attribute segments from existing Braille data, partially exemplified in Table 3. During the data augmentation phase, new Braille sequences are synthesized by leveraging the dependency syntax trees of original Braille sequences and replacing constituent segments with semantically compatible counterparts from the knowledge base. Segment similarity matching algorithm is employed to ensure the syntactic coherence of the reconstructed sequences.

4 Methodology

This study presents a fine-tuning framework for the Braille domain based on large-scale language models (LLMs), which integrates Braille-specific prior knowledge into the fine-tuning process to address the limitations of existing LLMs in handling Braille tasks, as illustrated in Fig. 4. Our approach introduces two distinct methods, CBKFT and EBKFT, tailored for Chinese and English Braille respectively, leveraging linguistic characteristics of each language. The unified fine-tuning procedure combines these methods to enhance the Braille comprehension and generation capabilities of the model.

4.1 CBKFT: Fine-Tuning Based on Chinese Braille Prior Knowledge

Chinese Braille is a syllable-based system, where each Braille fragment may correspond to multiple Chinese characters, but typically represents a single pinyin syllable (e.g., HV2 → hàn). Since LLMs lack inherent understanding of Braille symbols, CBKFT bridges this gap by mapping Braille fragments to

[3] http://www.braille.org.cn/

[4] https://github.com/liblouis/liblouis

Braille task type	Source sequence	Target sequence
Chinese-to-Braille	经济的快速发展	G*AGI D KYSU F9/V’
Pinyin-to-Braille	Jing1ji4defa1zhan3	G*AGI D KYSU F9/V’
Formula-to-Braille	$frac14x=15$	#A4;X 7#AE
Braille word segmentation	G*AGIDKYSU F9/V’	G*AGI D KYSU F9/V’
Mixed Chinese to Braille	故答案为：y	GU D91V W- #Y
Mixed English to Braille	Suppose that $g(x) = 5x - 3$.	,SUPPOSE T _ % G(X) .K #5X-3

Table 2: Examples of some Braille task data.

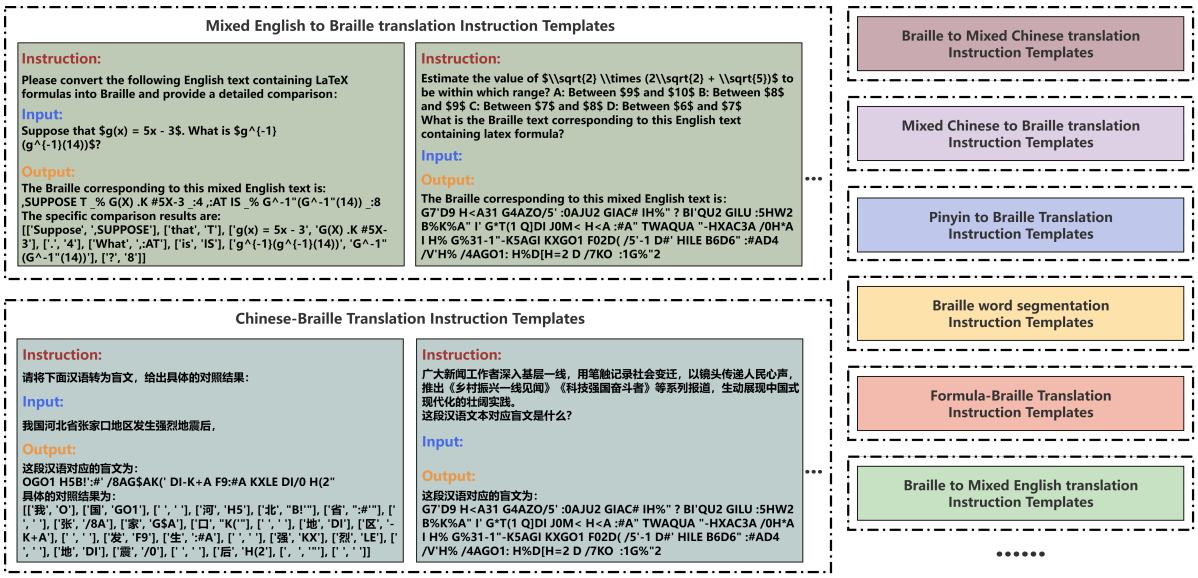

Figure 2: Instruction template samples.

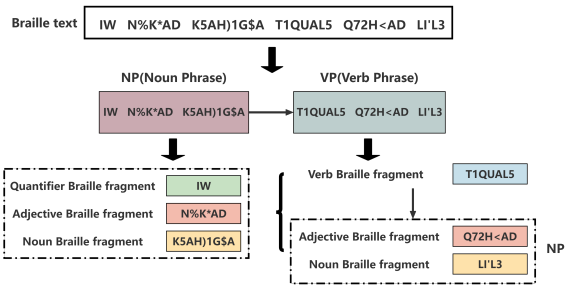

Figure 3: Example of component analysis of Braille text.

their phonetic syllables. We construct a *Chinese Braille Prior Knowledge Base* $\mathcal{K}_C = \{(b_i, p_j)\}$, where $b_i \in \mathcal{B}_C$ denotes a Braille fragment and $p_j \in \mathcal{P}$ represents its corresponding Pinyin syllable. Here, $\mathcal{B}_C$ is the set of 1,152 Chinese Braille fragments, and $\mathcal{P}$ is the Pinyin syllable inventory.

Given an input Braille sequence $S_C = \{b_1, b_2, \ldots, b_n\}$, we tokenize S_C into Braille-specific tokens (e.g., <|G*A|>, <|GI|>, <|F9|>, <|/V'|>) and map each token to its syllable p_j via $\mathcal{K}_C$. Let $\mathcal{T}_p \subseteq \mathcal{V}$ denote the subset of token IDs in the LLM's vocabulary that correspond to syllable p_j, where $\mathcal{V}$ is the full vocabulary. The embedding of Braille token b_i, denoted $\mathbf{e}_{b_i}$, is initialized as the mean of embeddings for all tokens in $\mathcal{T}_p$:

$$\mathbf{e}_{b_i} = \frac{1}{|\mathcal{T}_p|} \sum_{t \in \mathcal{T}_p} \mathbf{E}_t, \tag{1}$$

where $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times d}$ is the LLM's embedding matrix, and d is the embedding dimension. This syllable-aware initialization injects phonetic priors into Braille token embeddings, enabling the model to leverage shared phonological features across homophonic characters during fine-tuning. Formally: $\mathbf{E}[b_i] \leftarrow \mathbf{e}_{b_i}$.

4.2 EBKFT: Fine-Tuning Based on English Braille Prior Knowledge

English Braille fragments map directly to whole words rather than syllables (e.g., L1/ → least). EBKFT utilizes an *English Braille Prior Knowl-*

Attributes of the Braille fragments	Number of fragments	Fragment examples
Verb Braille fragment	11048	%'-QUA | T(1SU
Personal name Braille fragment	6717	K4'K*2H5 | :AD51-:A
Quantifier Braille fragment	3658	B9A:1SVASW | SVAF0/AR2
Place name Braille fragment	3482	H<AHIALV1 | TU'KUMV2SATV'

Table 3: Example of partial Braille fragments attribute knowledge base.

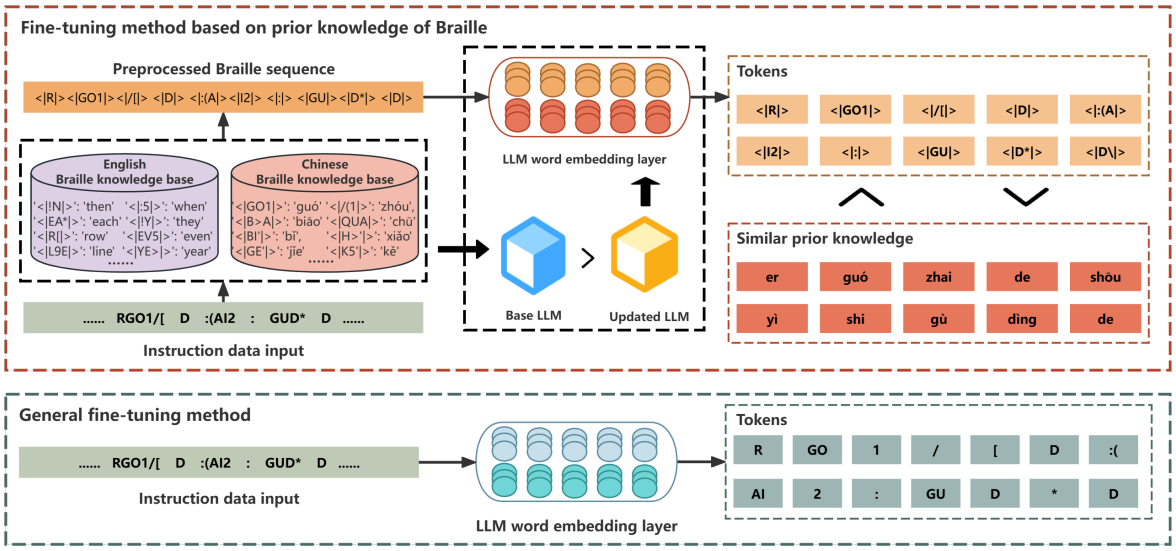

Figure 4: Comparison of fine-tuning based on Braille prior knowledge with conventional approaches. This study constructs a mapping knowledge base between Chinese/English Braille segments and their semantically equivalent words, enabling more rational segmentation of Braille sequences. Simultaneously, Braille segment embeddings are initialized with high-frequency equivalent word embeddings, which alleviates the model's comprehension difficulty of Braille patterns during fine-tuning.

edge Base $\mathcal{K}_E = \{(b'_i, w_j)\}$, where $b'_i \in \mathcal{B}_E$ is an English Braille fragment and $w_j \in \mathcal{W}$ is its corresponding word, with $\mathcal{B}_E$ containing 6,000 common English Braille fragments. For an input Braille sequence $S_E = \{b'_1, b'_2, \ldots, b'_m\}$, each token (e.g., <|L1/|>) is mapped to its word w_j via $\mathcal{K}_E$. The embedding of b'_i, $\mathbf{e}_{b'_i}$, is initialized by cloning the embedding of w_j: $\mathbf{E}[b'_i] \leftarrow \mathbf{E}[w_j]$, where $\mathbf{E}[w_j]$ is the pretrained embedding of word w_j. This direct lexical transfer preserves semantic and syntactic knowledge from the LLM's original vocabulary.

4.3 Model Fine-Tuning Procedure

The fine-tuning procedure unifies the Braille-specific initialization strategies of CBKFT and EBKFT with end-to-end training to adapt the LLM for Braille understanding and generation. Given a parallel dataset $\mathcal{D} = \mathcal{D}_C \cup \mathcal{D}_E$ containing Braille sequences paired with their textual counterparts, the training objective minimizes the negative log-likelihood of generating target text $Y = \{y_1, \ldots, y_T\}$ conditioned on Braille input S. The procedure involves three key phases: Braille tokenization and embedding initialization, contextual adaptation, and cross-lingual joint optimization.

For each Braille sequence S, we first apply language-specific tokenization rules to segment it into Braille fragments (e.g., <|G*A|> for Chinese or <|L1/|> for English). Each fragment b_i is mapped to its linguistic counterpart (Pinyin syllable p_j or word w_j) via the prior knowledge base $\mathcal{K}_C$ or $\mathcal{K}_E$. The token embedding $\mathbf{e}_{b_i}$ is initialized using the procedure described for CBKFT or EBKFT, ensuring $\mathbf{E}[b_i]$ encodes syllable-level or word-level semantics before training.

During training, the model processes the initialized Braille embeddings alongside standard text tokens, enabling it to learn contextual relationships between Braille fragments and their textual meanings. The embeddings of Braille tokens are updated

dynamically through backpropagation, guided by the loss $\mathcal{L}(\theta)$. To enhance cross-lingual generalization, we interleave batches from $\mathcal{D}_C$ and $\mathcal{D}_E$, forcing the model to distinguish between Chinese and English Braille patterns while sharing underlying linguistic knowledge. This phased approach ensures that the model first grounds Braille tokens in linguistic priors, then refines their representations through task-specific contextual learning, ultimately achieving robust performance on low-resource Braille tasks.

5 Experiment

5.1 Implementation Details

In this study, we used the pre-trained large language model Qwen2.5 7B as the base model, and proposed CBKFT and EBKFT fine-tuning methods based on Braille prior knowledge for Chinese and English Braille tasks, respectively. Comparable baseline models including LLama3 8B, GLM4 9B, and Qwen2.5 7B of similar scales were fine-tuned for performance comparison. All models were fine-tuned using LoRA adapters for efficient parameter updating, with a rank of 8, and the AdamW optimizer was used for training. The initial learning rate was set to 1e-4, and the maximum sequence length was fixed at 1024. To evaluate model performance on Braille domain tasks, we employed multiple metrics: sacreBLEU [5] and chrF++ for translation quality assessment, CER (Character Error Rate) and TER (Translation Error Rate) for error rate analysis, along with BERTScore (Zhang et al., 2019) for semantic similarity evaluation between the model's generated answers and the reference answers. All experiments were conducted on a computing platform equipped with two NVIDIA A40 GPUs.

5.2 Performance of BKFT

The fine-tuning method based on Braille prior knowledge enables large language models to comprehend Braille sequences better, thereby enhancing Braille-to-text translation performance. Tables 4 present the performance of various models under different fine-tuning data sizes for Braille-to-Chinese translation tasks. BrailleLLM achieves the best performance across all data scales, attaining a CER of only 0.0542 with 10,000 fine-tuning samples, demonstrating low character error rates in the generated Chinese text. The results indicate that the performance improvement is more pronounced with smaller fine-tuning datasets. With only 3,000 fine-tuning samples, our method achieves an 18.96 BLEU score improvement compared to approaches without BKFT. This suggests that the BKFT method can effectively learn Braille sequence features even in low-resource scenarios. The comparison against the Qwen2.5-7B baseline, which undergoes only conventional fine-tuning, effectively ablates our BKFT method and validates its significant contribution to the performance improvement.

5.3 Braille Translation in Application Scenarios

Due to spatial perception challenges, visually impaired individuals often make Braille writing errors such as missing dots, extra dots, and other inaccuracies when embossing six-dot Braille characters. Consequently, Braille in real-world applications typically contains such errors and consists of mixed elements, including numbers, punctuation, and mathematical symbols. To further evaluate the effectiveness of the BKFT method, we integrated common writing errors into the fine-tuning data. As shown in Table 5, our approach consistently outperforms baseline models on the mixed English Braille-to-text conversion task with erroneous inputs. This demonstrates that BrailleLLM acquires a more robust understanding of imperfect Braille during fine-tuning, highlighting its practical applicability in real-world scenarios.

5.4 Auxiliary Performance of Braille Data Augmentation

To evaluate the effectiveness of Braille data augmentation method, we compared it with two commonly used data augmentation strategies: noise injection-based adversarial generation and segment replacement-based adversarial generation. The noise injection method randomly deleted or inserted 15% of corresponding Braille-text pairs, while the fragment-replacement method replaced fragments within Braille-text pairs. After augmenting 3,000 Braille-Chinese parallel fine-tuning samples, the experimental results are shown in Table 6. Our proposed syntax tree-based Braille augmentation method achieved the best auxiliary performance, demonstrating that our approach can effectively enhance the diversity of Braille data and provide more Braille-specific features for the fine-tuning process.

[5] https://github.com/mjpost/sacreBLEU

Model	sacreBLEU ↑			chrF++ ↑			CER ↓		
	3K	6k	10K	3K	6k	10K	3K	6k	10K
GLM4 9B	51.04	78.78	84.89	49.16	66.15	73.33	0.5670	0.1422	0.0932
Llama3 8B	52.77	70.29	77.48	42.21	58.50	68.07	0.3372	0.2250	0.0889
Qwen2.5-7B	63.19	77.95	85.52	50.39	65.63	74.60	0.2660	0.1509	0.0870
BrailleLLM	**82.15**	**88.37**	**91.15**	**69.69**	**77.82**	**82.18**	**0.1180**	**0.0739**	**0.0542**

Table 4: Performance of various models under different fine-tuning data quantities for Braille-to-Chinese translation tasks. Note: BrailleLLM (Qwen2.5 7B + BKFT).

Model	sacreBLEU ↑			chrF++ ↑			TER ↓		
	3K	6k	10K	3K	6k	10K	3K	6k	10K
GLM4 9B	77.74	86.00	87.03	76.39	82.91	86.90	20.04	11.60	12.77
Llama3 8B	78.80	84.90	88.21	76.51	81.62	85.85	18.51	14.72	10.63
Qwen2.5-7B	76.92	85.85	88.57	73.50	82.85	86.27	24.78	12.18	11.09
BrailleLLM	**81.85**	**87.52**	**89.64**	**79.19**	**84.60**	**87.39**	**16.87**	**10.31**	**9.95**

Table 5: Performance of different models in the applied scenario of mixed English Braille-to-plaintext translation.

Method	Chinese-to-Braille
None	79.49
Noise Injection	85.71
Fragment Replacement	83.19
Ours	**87.32**

Table 6: Auxiliary performance of different data augmentation methods.

5.5 Evaluation of BrailleLLM's Question-Answering Capability

For Braille-related tasks, user input formats are often inconsistent, requiring the model to comprehend the question before generating a response. To evaluate BrailleLLM's question-answering capability, we constructed 1,000 question-answer pairs covering various Braille tasks and assessed the semantic similarity between the model's responses and reference answers using BERTScore, as shown in Table 7. The results demonstrate that the model's answers exhibit high semantic alignment with the expected outcomes, proving its effectiveness in response generation and task processing.

Furthermore, we compared BrailleLLM's performance on Braille tasks with general-purpose conversational LLMs, including GPT-4 and Claude 3.5. Current general models exhibit limited capability in processing Chinese Braille. Table 8 displays model performance on mixed English Braille-to-plaintext translation tasks. BrailleLLM demonstrates a significant performance advantage over existing conversational models in Braille-related tasks.

Method	Text-to-Braille	Braille-to-Text
English QA	0.9605	0.9543
Chinese QA	0.9487	0.9435

Table 7: Performance of BrailleLLM on question answering tasks.

Model	sacreBLEU	chrF++
GPT-4	48.14	38.69
Claude3.5	22.76	18.50
BrailleLLM	**89.94**	**87.39**

Table 8: Comparative performance of BrailleLLM versus general-purpose LLMs on Braille-specific tasks.

6 Conclusion

The paper proposes BrailleLLM, the first multi-task large language model for Braille processing, which enables bidirectional conversion between Braille and mixed-content texts containing formulas and regular text. We construct and open-source a large-scale annotated dataset covering Chinese and English Braille, mitigating the long-standing data scarcity issue in Braille research. An efficient fine-tuning approach tailored for Braille data is proposed to perform instruction tuning with LLMs across various Braille tasks. The fine-tuned model

demonstrates superior Braille translation performance and exhibits Braille processing capabilities absent in existing general-purpose conversational LLMs. This work not only enhances understanding of general-purpose models in the Braille domain but also lays the groundwork for future exploration.

Limitations

Although the proposed Braille dataset and fine-tuning method show significant effectiveness for Braille tasks, we acknowledge that Braille is often represented in image form, which limits the method's applicability. Future work could expand the dataset to include more forms of Braille input, ensuring better Braille information processing services.

Acknowledgements

This work is supported by the National Natural Science, Foundation of China (Grant No. 62372408).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Azizah Awang, Intan Farahana Abdul Rani, Kway Eng Hock, Nur Adillah Ramly, and Rozita Kamil. 2024. Innovative approaches in teaching early braille reading skills: A theory and practice study. *Journal of Contemporary Social Science and Education Studies (JOCSSES) E-ISSN-2785-8774*, 4(2):177–200.

Rachel Bawden, Nikolay Bogoychev, Ulrich Germann, Roman Grundkiewicz, Faheem Kirefu, Antonio Valerio Miceli Barone, and Alexandra Birch. 2019. The university of edinburgh's submissions to the wmt19 news translation task. *arXiv preprint arXiv:1907.05854*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jia Cai, Xiangdong Wang, Lizhen Tang, Xiaojuan Cui, Hong Liu, and Yueliang Qian. 2019. Automatic Chinese-Braille Conversion Based on Chinese-Braille Contrasted Corpus and Deep Learning (in Chinese). *Chinese Journal of Information*, 33(4):60–67.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mattia A Di Gangi and Marcello Federico. 2017. Monolingual embeddings for low resourced neural machine translation. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 97–104.

Jun Dong. 2023. Transfer learning-based neural machine translation for low-resource languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Christian Egli. 2009. Liblouis–a universal solution for braille transcription services. In *Proceedings of Daisy 2009 Conference*.

Yuan Gao, Feng Hou, and Ruili Wang. 2024. A novel two-step fine-tuning framework for transfer learning in low-resource neural machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3214–3224.

Junjie Hu, Hiroaki Hayashi, Kyunghyun Cho, and Graham Neubig. 2021. Deep: denoising entity pre-training for neural machine translation. *arXiv preprint arXiv:2111.07393*.

Tianyuan Huang, Wei Su, Lei Liu, Jing Zhang, Chuan Cai, Yongna Yuan, and Cunlu Xu. 2023. Translating braille into chinese based on improved cbhg model. *Displays*, 78:102445.

Nikisha B Jariwala and Bankim Patel. 2017. Conversion of 2d mathematical equation to linear form for transliterating into braille: an aid for visually impaired people. *Int JInnov Res Sci Eng Technol*, 6(4).

Minghu Jiang, Xiaoyan Zhu, Georges Gielen, Elliott Drábek, Ying Xia, Gang Tan, and Ta Bao. 2002. Braille to print translations for chinese. *Information and software Technology*, 44(2):91–100.

Qi Jiang, Wei Su, Ying Xie, Anping Zhouhong, Jiuwen Zhang, and Chuan Cai. 2021. End-to-End Chinese-Braille automatic conversion based on Transformer (in Chinese). *Computer Science*, 48(11A):136.

CVRT Johan Rempel and CPACC CATIS. 2022. The importance of braille during a pandemic and beyond. *Assistive Technology Outcomes & Benefits*, 16(2):127–134.

Fituma Yadasa Kana and Asmerom Tekle Hagos. 2024. Factors hindering the use of braille for instruction and assessment of students with visual impairments: A systematic review. *British Journal of Visual Impairment*, page 02646196241239173.

Fuxue Li, Beibei Liu, Hong Yan, Mingzhi Shao, Peijun Xie, Jiarui Li, and Chuncheng Chi. 2024. A bilingual templates data augmentation method for low-resource neural machine translation. In *International Conference on Intelligent Computing*, pages 40–51. Springer.

Rumeng Li, Xun Wang, and Hong Yu. 2020. Metamt, a meta learning method leveraging multiple domain data for low resource machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8245–8252.

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

Yu Lu, Jiali Zeng, Jiajun Zhang, Shuangzhi Wu, and Mu Li. 2022. Learning confidence for transformer-based neural machine translation. *arXiv preprint arXiv:2203.11413*.

Agustín Lucas, Alexis Baladón, Victoria Pardiñas, Marvin Agüero-Torales, Santiago Góngora, and Luis Chiruzzo. 2024. Grammar-based data augmentation for low-resource languages: The case of guarani-spanish neural machine translation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6385–6397.

Jiang Minghu, Zhu Xiaoyan, Xia Ying, Tan Gang, Yuan BaoZong, and Tang Xiaofang. 2000. Segmentation of mandarin braille word and braille translation based on multi-knowledge. In *WCC 2000-ICSP 2000. 2000 5th International Conference on Signal Processing Proceedings. 16th World Computer Congress 2000*, volume 3, pages 2070–2073. IEEE.

Cheonbok Park, Yunwon Tae, Taehee Kim, Soyoung Yang, Mohammad Azam Khan, Eunjeong Park, and Jaegul Choo. 2020. Unsupervised neural machine translation for low-resource domains via meta-learning. *arXiv preprint arXiv:2010.09046*.

Ye Qi, Devendra Singh Sachan, Matthieu Felix, Sarguna Janani Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? *arXiv preprint arXiv:1804.06323*.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Felipe Sánchez-Martínez, Víctor M Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Mikel L Forcada, Miquel Espla-Gomis, Andrew Secker, Susie Coleman, and Julie Wall. 2020. An english-swahili parallel corpus and its use for neural machine translation in the news domain. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 299–308.

Rico Sennrich. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. Lawllm: Law large language model for the us legal system. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4882–4889.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Xiangdong Wang, Yang Yang, Hong Liu, and Yueliang Qian. 2016. Chinese-braille translation based on braille corpus. *International Journal of Advanced Pervasive and Ubiquitous Computing (IJAPUC)*, 8(2):56–63.

Xiangdong Wang, Yang Yang, Jinchao Zhang, Wenbin Jiang, Hong Liu, and Yueliang Qian. 2017. Chinese to braille translation based on braille word segmentation using statistical model. *Journal of Shanghai Jiaotong University (Science)*, 22:82–86.

Xiangdong Wang, Jinghua Zhong, Jia Cai, Hong Liu, and Yueliang Qian. 2019. Cbconv: service for automatic conversion of chinese characters into braille with high accuracy. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 566–568.

HaiLong Yu, Wei Su, Lei Liu, Jing Zhang, Chuan Cai, and Cunlu Xu. 2023. Pre-training model for low-resource chinese–braille translation. *Displays*, 79:102506.

RH Zatserkovnyi, VZ Mayik, RS Zatserkovna, and L Ya Mayik. 2019. Analysis of braille translation software. *Printing and Publishing*, 2:36–44.

Ju-Xiao Zhang, Hai-Feng Chen, Bing Chen, Bei-Qin Chen, Jing-Hua Zhong, and Xiao-Qin Zeng. 2022. Design and implementation of chinese common braille translation system integrating braille word segmentation and concatenation rules. *Computational Intelligence and Neuroscience*, 2022(1):8934241.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Chulun Zhou, Fandong Meng, Jie Zhou, Min Zhang, Hongji Wang, and Jinsong Su. 2022. Confidence based bidirectional global context aware training framework for neural machine translation. *arXiv preprint arXiv:2202.13663*.

A Appendix

A.1 Data Validation Pipeline

To ensure the fidelity of our bilingual Braille dataset, we designed a rigorous validation pipeline executed by Braille experts. This process systematically verifies correctness across five stages, from low-level format integrity to high-level semantic coherence. Table 9 outlines the key aspects of each stage.

Stage	Key Aspects
Automated Check	Checks for invalid Braille ASCII codes and verifies standard LaTeX formatting.
Unit Accuracy	Verifies the correct mapping of individual plaintext elements (e.g., Chinese characters, words, formulas) to Braille, including spelling and tone mark accuracy.
Rule Validation	Ensures adherence to Braille language rules, such as correct Braille word segmentation, abbreviated tone marks, and contractions.
Cross-Review	Involves independent review by a different expert on a 20% data sample to ensure high Inter-Annotator Agreement.
Fluency Review	Involves a tactile read-through of the Braille to check for natural transitions between text and formulas, and to ensure semantic coherence and absence of contextual ambiguity.

Table 9: The Five-Stage Data Validation Pipeline.

A.2 Qualitative Analysis of Error Patterns

To gain deeper insights into the superior performance of BrailleLLM on complex Braille translation tasks, we conduct a qualitative analysis of its error patterns. The core challenge in Braille translation stems from its inherent ambiguity: a single Braille cell or sequence can map to diverse elements, such as characters from different languages, digits, punctuation, or mathematical symbols. This polysemy necessitates a strong contextual understanding of Braille. As illustrated in Table 10, the baseline model—a traditionally fine-tuned Qwen2.5-7B—exhibits typical error patterns arising from this ambiguity due to its lack of sufficient Braille-specific prior knowledge. For instance, in Case 1, the baseline model incorrectly translates Chinese Braille into the English word "Nine" and makes errors based on phonetic similarity. In Case 2, it introduces a spurious "⇒" at the junction of a formula and text, while also omitting characters. In contrast, BrailleLLM, with its inherent Braille knowledge, effectively mitigates such ambiguity during the translation process.

Case 1: Plain text Braille translation	
Prompt	Please translate the following Braille into plain text: `M831 &1+1 #AI:GI D F9'GO1" \\1 N%\\2 :1 <A I2Y :AM* D ,LOUIS ,BRAILLE SO' F9M*"2`
Reference	盲文源于19世纪的法国，由年幼时因意外失明的Louis Braille所发明。 Braille originated in 19th-century France and was invented by Louis Braille, who lost his sight in a childhood accident.
Baseline	盲文源于19世纪的法国，由Nine时因意外声明的LOUIS BRAILLE所发明。
BrailleLLM	盲文源于19世纪的法国，由年幼时因意外失明的LOUIS BRAILLE所发明。
Case 2: Mixed text Braille translation	
Prompt	Please translate the following mixed Braille text into plain text: `GUH>'M*@ H>'G_ZU' HO2D51 ^:01S] H>'ZU'^G[L+ D ZWD9/1 W#E16 \" C':1 ;P*1 7;P*2 7#A2 4`
Reference	求解变量 y 的值，并注意定义域：`$\frac{3}{y-2} = \frac{5}{y+4}$` Solve for the variable y, and pay attention to the domain: `$\frac{3}{y-2} = \frac{5}{y+4}$`
Baseline	求解变量 y 的值，凭注意定义：⇒`$\frac{3}{y-2}=\frac{5}{y+4}$`
BrailleLLM	求解变量 y 的值，凭 注意定义域：`$\frac{3}{y-2}=\frac{5}{y+4}$`

Table 10: Case study of Braille translation. The baseline model is Qwen2.5-7B with conventional fine-tuning. Translation errors are highlighted in red.