Vision-Braille: A Curriculum Learning Toolkit and Braille-Chinese Corpus for Braille Translation

Alan Wu[2]*, Ye Yuan[1], Zhiping Xiao[3], Ming Zhang[1]

[1]State Key Laboratory for Multimedia Information Processing, School of Computer Science, PKU-Anker LLM Lab, Peking University
[2]College of Agriculture and Life Sciences, Cornell University
[3]Paul G. Allen School of Computer Science and Engineering, University of Washington
Correspondence: Ming Zhang (mzhang_cs@pku.edu.cn)

Abstract

We present Vision-Braille, the first publicly available end-to-end system for translating Chinese Braille extracted from images into written Chinese. This system addresses the unique challenges of limited annotated resources and tone omission. The task is out of reach for general-purpose LLMs: given the complete Chinese Braille encoding specification and a worked example in context, Qwen3-235B-A22B reaches only 9.46 BLEU, and the only prior Braille-to-Chinese model with released checkpoints scores 5.36 BLEU on passage-level input. Braille is effectively absent from pretraining corpora, decoding it demands exact multi-step symbolic reasoning, and writers omit up to 90% of tone markers, leaving homophones that must be recovered from context. Vision-Braille integrates a Braille OCR pipeline with an open-weights LLM fine-tuned on a synthetic Braille-Chinese corpus that reproduces authentic tone-omission habits, using a four-stage curriculum that moves from short, fully toned input to long passages with tone markers almost entirely stripped. On passage-level translation with 10% tone retention, Curriculum-Braille achieves 83.28 BLEU at a 9.94% character error rate, and 76.66 BLEU on passages hand-transcribed by an expert from the Chinese Braille Library, with no reliance on any proprietary API. For the 150,000 school-age blind children in China, Vision-Braille offers an inclusive NLP solution that supports participation in mainstream education by enabling teachers to grade Braille homework without extensive training.[1]

1 Introduction

Among the 4 million registered people with visual disabilities in China (China Disabled Persons' Federation, 2022b), there is a special group of 150,000 school-age blind children (Li et al., 2021) who face the risk of "double-blind" (physically disabled and illiterate). We have found that even though the population is staggering, education resources are scarce: only 11 Senior High Schools for Blind Students exist nationwide, all of them in major cities, and together they enrolled just 443 new blind students in 2022 (China Disabled Persons' Federation, 2022a). Although the Chinese Bureau of Education enacted the specialized form of GaoKao, allowing participants to read Braille exam papers and write in Braille, this group was almost absent in higher education: in 2025 only 16 students nationwide sat this special GaoKao, against a total of 13.35 million GaoKao candidates (Ministry of Education of the People's Republic of China, 2025). The scarcity of specialized educational resources has discouraged many eager learners from pursuing their college dreams. However, if artificial intelligence could be leveraged to support these "double-blind" children, it might enable them to participate in mainstream education and transform their lives.

The Braille writing system, devised by Louis Braille in 1824, is a tactile reading and writing method used worldwide by individuals with visual impairments. A Braille cell is uniform worldwide, consisting of six raised dots. Braille has been adapted to represent each Chinese character phonetically through up to three cells: consonants, vowels, and tones; words with multiple Braille characters are separated by spaces based on their meaning. For ease of writing, Chinese Braille omits consonants, vowels, or tones based on rules and personal preferences, particularly tones. According to research on the largest Chinese Braille corpus, writers of Current Braille (现行盲文) on average omit up to 90 percent of tones (Xiao et al., 2020). While this aids writing, it obstructs accurate understanding due to homonyms and characters sharing the same pronunciation. One example is the Chinese

*Work done during the remote internship at Dlib Group, School of Computer Science, Peking University.

[1]Our code, together with links to the data and model weights, is available at https://github.com/AlanYWu/VisionBraille.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 59658–59672
October 24-29, 2026 ©2026 Association for Computational Linguistics

pronunciation "yi," which maps to more than a hundred common Chinese characters, as shown in Figure 1. This issue also arises in French (Bouillon et al., 2013), Khmer (Huor et al., 2007), and Japanese (Komeiji and Tanaka, 2019).

We have interviewed several faculty members at the Beijing School for Blind Children and have observed the need for more adequate educational resources. In this study, we design the "Vision-Braille" system, providing an end-to-end Braille recognition and translation pipeline as an educational tool for visually impaired Chinese students. The overall pipeline is shown in Figure 1; it consists of Braille OCR (Image-to-Braille) and Braille Translation (Braille-to-Text), described below.

2 Related Work

2.1 Braille Translation

Machine translation has evolved from rule-based and statistical methods through RNN and LSTM (Hochreiter and Schmidhuber, 1997) architectures to transformer-based models (Vaswani et al., 2017), with recent large language models such as Qwen3 (Yang et al., 2025) showing promise in low-resource and instruction-tuned translation settings.

This progression is mirrored in Braille-to-Chinese translation, where early systems relied on rule-based or N-gram models (Wang et al., 2010; Jiang et al., 2002), followed by neural approaches using LSTM and GRU (Wang, 2019). Yu et al. (2023) proposed a Transformer-based model with Braille-specific pretraining objectives, achieving strong BLEU scores through MacBERT-based encoders. Huang et al. (2024) jointly trained bidirectional Braille-Chinese models and distilled them into a unified decoder architecture. Most recently, BrailleLLM (Huang et al., 2025) applies instruction tuning to unify Braille translation, formula conversion, and mixed-text tasks, using Braille Knowledge-Based Fine-Tuning (BKFT) to initialize character-level symbol understanding.

However, these studies primarily focus on sentence-level data, lacking the longer-range dependencies found in real-world Braille passages. In contrast, our work moves beyond sentence-level translation by incorporating explicit segmentation signals, applying curriculum learning to transition from short to long-form passages, and evaluating on an open-domain test suite derived from student-generated Braille.

2.2 Low-Resource Translation and Curriculum Learning

Chinese Braille is a low-resource language: although closely related to Chinese, it encodes pronunciation via pinyin rather than characters, and standard Chinese corpora do not record pronunciation information. Furthermore, the only known Braille–Pinyin–Chinese corpus is not publicly available (Xiao et al., 2020). Previous work has explored unsupervised and few-shot methods (Nguyen et al., 2024) as well as in-context learning (Cahyawijaya et al., 2024) for low-resource languages.

Curriculum learning (CL) offers a complementary strategy by presenting examples in an easy-to-hard order, yielding faster convergence and better minima (Bengio et al., 2009). Our approach adopts this principle: given that Chinese-Braille corpora are limited and prone to transcription confusion, we design a curriculum that begins with short sentences and gradually introduces longer passages, smoothing the learning signal and enabling LLMs to generalize reliably.

2.3 Braille Recognition

A complete Braille translation system must also convert physical Braille images into symbolic representations. Several CNN-based detection models (Li et al., 2018; Ovodov, 2021; Li et al., 2020) have achieved satisfactory accuracy on datasets such as DSBI and Angelina. We adopt the Angelina Braille Reader (Ovodov, 2021), a RetinaNet-based detector with strong real-world performance on Braille recognition tasks.

3 Dataset

3.1 Dataset Construction

We describe how the parallel corpus is built, covering data collection and post-processing.

Collection. Braille-to-Chinese parallel datasets are extremely scarce and all private, owing to the labor-intensive nature of double-checking and alignment processes that require specialists proficient in both Braille and Chinese (Xiao et al., 2020). In contrast, the conversion from Chinese to Braille follows systematic linguistic rules and can be automated. To overcome these limitations, we automatically generate synthetic parallel datasets by converting Chinese corpora into Braille using rule-based tools. We construct two paired Chinese-Braille datasets: one at the sentence level and one at the passage level.

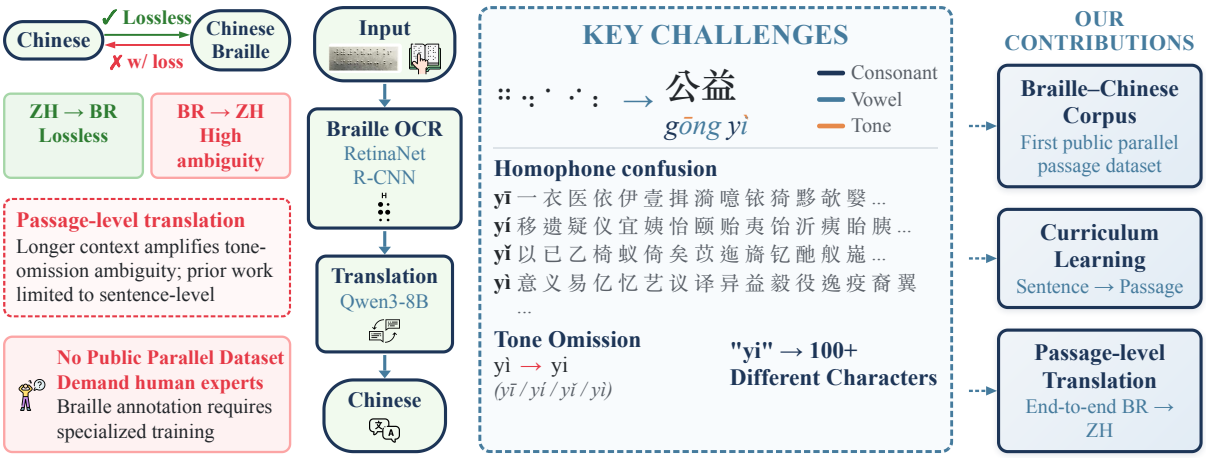

Figure 1: The pipeline of our Curriculum-Braille.

The sentence-level Chinese corpus was sourced from the publicly available Leipzig Corpora Collection (Goldhahn et al., 2012), comprising one million sentences collected from news media between 2007 and 2009. These sentences were converted into Braille using the conversion tool provided by the Digital Platform for Chinese Braille (Zhongguo Mangwen Shuzi Pingtai [Digital Platform for Chinese Braille], 2019). This system annotates Chinese character pronunciations and converts them into "fully toned" Braille. For the passage-level dataset, we curated aligned Braille and Chinese text pairs from publicly available Braille book resources hosted on The Braille Online Platform of China. All data pairs underwent preprocessing to remove unrecognized tokens and were truncated to a maximum of 2048 tokens.

Processing. To simulate tone omission in real-world Braille input, we apply the tone-deletion operator defined in §4.2.3: each tone marker is removed *independently at random* (with a fixed seed for reproducibility), governed by a retention rate r. We deliberately use unstructured random deletion rather than hand-coding the 省写 conventions of the national standard GF 0019–2018 (国家通用盲文方案). At low retention, random deletion drops tones indiscriminately, without exploiting the position-specific regularities that experienced writers follow, so the model must recover tone purely from context rather than from a fixed omission pattern. This is the harder setting, and it transfers to real usage: on expert-produced Braille written under standard 省写 conventions—i.e. genuine systematic tone loss—our model translates accurately (§5.5), despite never being trained on rule-based omission. Additionally, we remove all space characters from Braille sequences before training. Chinese is written without inter-word spaces, and in standard OCR pipelines space tokens are both unreliable and often dropped. Each Chinese character in Mandarin Braille typically begins with weakly distinctive and consistent dot patterns (e.g. initial consonant or vowel-based Braille clusters), so the model can implicitly learn word segmentation via these local cues without explicit space delimiters.

Level	# Train	Avg. Braille Len.	Avg. Chinese Len.
Sentence	376,846	116	31
Passage	17,592	963	332

Table 1: Dataset summary. Lengths are average token counts per sample in the training set.

3.2 Dataset Statistics

The collected data does not contain any personally identifiable information or offensive content. Our dataset is mainly built upon instances from real-world exam and book data. Therefore, it is unlikely to contain sensitive data. Finally, all datasets were split into training, validation, and test sets in a ratio of 8:1:1. Table 1 summarizes the key statistics. The passage-level pairs are roughly 8× longer than sentence-level pairs, presenting a substantially harder translation task. Per-split breakdowns are in Appendix C.

4 Methodology

4.1 Braille OCR

In our research, we employed RetinaNet for the task of Braille Optical Character Recognition (OCR). RetinaNet utilizes a pre-trained convolu-

tional neural network architecture, augmented by a Feature Pyramid Network (FPN) as its backbone, which effectively extracts multi-scale features from input images. This model facilitates one-stage object detection by integrating a classification subnet and a box regression subnet. The classification subnet, comprising several 3×3 convolutional layers and a sigmoid activation layer, predicts the presence of objects at specific spatial locations by generating binary outputs. Concurrently, the box regression subnet estimates the offset from each anchor to the actual bounding boxes. Subsequently, the identified regions are processed using a pre-trained image classification model to ensure precise object recognition. The incorporation of focal loss further enhances the model's accuracy, making it highly effective for one-stage object detection tasks.

However, the accuracy of Braille OCR algorithms is inherently constrained by factors such as lighting conditions, interference from Braille dots on the reverse side of the paper, and physical deformation of the paper itself. To quantify these limitations, we evaluated our Braille OCR model on 500 hand-written Braille cells from a single-sided page, observing a character-level error rate below 1%. This error rate rises to around 15% when interference from dots on the opposite side of the paper is present. While double-sided Braille printing is practical for tactile reading, visual recognition systems are notably affected by such interference, highlighting a critical consideration in Braille OCR applications.

Despite these OCR limitations, our Braille translation model aims to mitigate the impact of noise and errors introduced during the OCR process. Further details on our approach are provided in the following section on Braille Translation.

4.2 Braille Translation

We fine-tune a pre-trained Qwen3-8B large language model to improve our Braille-to-Chinese translation system. Qwen3 follows a Transformer architecture and is trained on large-scale bilingual Chinese and English corpora, with particularly strong coverage of Chinese, making it well-suited for our task. The model conditions on sequences of Braille tokens and autoregressively generates the corresponding Chinese output. By fine-tuning the model on our Braille dataset, we enable it to learn accurate alignments between Braille patterns and their corresponding Chinese characters. We select the 8B parameter variant to balance model capacity with computational cost.

We add 63 braille characters as special tokens into the vocabulary of the Qwen3-8B tokenizer and resize the embedding dimension of the Qwen3-8B model accordingly. Our fine-tuning task aims to teach the Qwen3-8B model to learn the rules and mappings of unseen braille tokens and accurately return the translated Chinese characters. We adopt the BLEU metric, specifically BLEU-4, for model evaluation. BLEU (Bilingual Evaluation Understudy) assesses the quality of machine translation by comparing the model's output to one or more reference translations, based on the precision of n-grams (contiguous sequences of words). BLEU-4 considers up to 4-gram matches, capturing both local word choice and short-range fluency. The score ranges from 0 to 100, with higher values indicating closer alignment with the reference and thus better translation quality.

4.2.1 Task Definition

We study Braille-to-Chinese (BR2ZH) translation. Given a sequence of Chinese Braille symbols, the model outputs a sequence of Chinese characters. We focus on robustness to missing tone markers in the Braille input, which increases ambiguity and requires contextual disambiguation.

4.2.2 Model and Prompting

We fine-tune QWEN3-8B with full supervised fine-tuning (SFT). We extend the tokenizer with a Braille vocabulary of 63 symbols. We do not introduce any special direction tokens. Instead, we specify the task direction in the system prompt (e.g., "Translate Chinese Braille to Chinese characters.").

Zero-shot prompting baseline. To establish a lower bound on what general-purpose LLMs can achieve without any Braille-specific training, we evaluate Qwen3-8B and Qwen3-235B-A22B in a zero-shot prompting setting. We inject the core Chinese Braille translation rules (including the Braille symbol inventory, pinyin-to-Braille mappings, and tone marker conventions) into the system prompt, and then provide the raw Braille input for the model to translate. No fine-tuning or in-context examples are used; the model must rely entirely on its pretrained knowledge combined with the rule descriptions provided in the prompt. This baseline tests whether LLMs can perform Braille translation from rule specifications alone, without exposure to parallel training data.

4.2.3 Tone Marker Deletion Operator

Let $\mathcal{V}_{\text{br}}$ be the Braille symbol vocabulary and $\mathcal{T} \subset \mathcal{V}_{\text{br}}$ be the subset of symbols that encode tones. We define a tone retention rate $r \in [0, 100]$. Given an input Braille sequence $\mathbf{x}$, we construct $\tilde{\mathbf{x}} = \mathrm{DropTone}(\mathbf{x}; r)$ by independently deleting each occurrence of a tone marker $t \in \mathcal{T}$ with probability $1 - r/100$, while keeping all non-tone symbols unchanged. We use a fixed random seed so that $\mathrm{DropTone}$ is deterministic and reproducible for a given example.

4.2.4 Four-stage Training Curriculum

We train the final BR2ZH model via four stages (Figure 2). Stages 1–2 each run for one epoch; Stage 3 consists of 10 sequential sub-stages of one epoch each; Stage 4 runs for three epochs. We reset the cosine learning rate schedule at the beginning of each stage and sub-stage. All stages are trained on 8 GPUs. We set the maximum sequence length to 2048, and no training example exceeds this limit in our latest preprocessing pipeline.

Stage 1: Foundation (sentence-level, full tones). We fine-tune on the full BR2ZH sentence-level training set. We preserve all tone markers in the input Braille ($r = 100$).

Stage 2: Long-range (passage-level, full tones). We fine-tune on the full BR2ZH passage-level training set. We preserve all tone markers in the input Braille ($r = 100$).

Stage 3: Tone curriculum (passages, 10 sub-stages, scheduled retention). We perform a tone-robustness curriculum on passage-level data, training sequentially through 10 sub-stages with decreasing tone retention. For sub-stage $k \in \{1, \ldots, 10\}$, we set

$$r_k = 110 - 10k \in \{100, 90, 80, \ldots, 10\}, \quad (1)$$

and apply $\mathrm{DropTone}(\cdot; r_k)$ to all Braille inputs in that sub-stage. Each sub-stage trains for one full epoch on the passage-level data at its assigned retention rate, loading from the previous sub-stage's checkpoint and resetting the learning rate schedule.

Stage 4: Consolidation (passage-level, low tone, multi-epoch). We perform a final consolidation stage on passage-level data with a fixed tone retention rate of $r = 10$, training for three epochs. This extended training allows the model to solidify its ability to translate passages under realistic tone-omission conditions, where the majority of tone markers are absent. By repeatedly exposing the model to low-tone passage data, Stage 4 reinforces the robustness learned in Stage 3 (Tone curriculum) and improves convergence on long-form translation.

4.2.5 Inference

For all reported results, model predictions are generated autoregressively during evaluation with top-p sampling (temperature 0.95, p = 0.7, top-k 50). We evaluate generated outputs using BLEU-4, following standard machine translation practice.

5 Experiments

We fine-tune Qwen3-8B-Instruct via full supervised fine-tuning on 8× NVIDIA H20 GPUs with DeepSpeed ZeRO-3 and bf16 precision. Training details and hyperparameters for each stage are provided in Appendix B (Table 5).

We have constructed a complete Braille translation pipeline incorporating Braille recognition and Braille-to-Chinese translation. Users scan Braille images using their phones; the images are uploaded to a remote server and processed by the RetinaNet-based Braille recognition system. The recognized digital Braille characters are then translated into Chinese by our model. Finally, the raw translation is polished by the Gemini-3.1-Pro API with Low reasoning effort to correct homophone substitutions and punctuation errors (see Appendix E).

5.1 Curriculum Learning for Learning Tone Omission

We adopt a curriculum learning approach specifically designed to enhance the model's robustness to tone omission. Stage 3 (Tone curriculum) of our four-stage curriculum (§4.2.4) implements this by training sequentially through 10 sub-stages on passage-level data, where the tone retention rate r decreases from 100% to 10% in steps of 10. Each sub-stage trains for one full epoch, loading from the previous sub-stage's checkpoint and resetting the learning rate schedule. This progressive exposure enables the model to incrementally recognize implicit tonal cues from context as explicit tone markers grow sparse.

Stage 4 (Consolidation) then consolidates this robustness by training for three additional epochs on passage-level data with a fixed tone retention rate of $r = 10$, simulating realistic conditions where the majority of tone markers are absent. By structuring tone-omission training in this manner, we cultivate

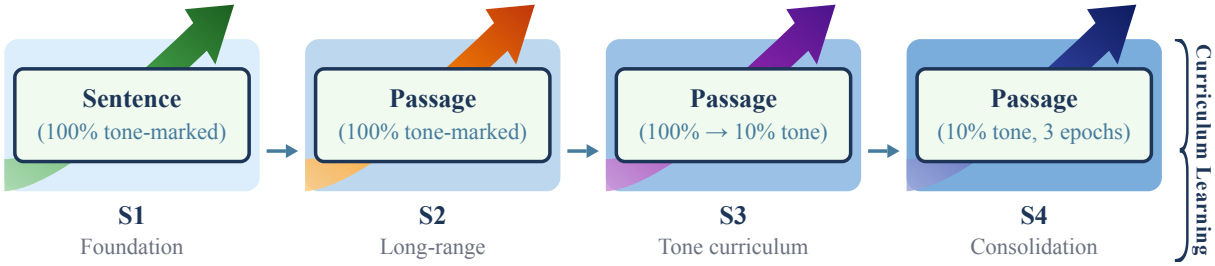

Figure 2: Overview of the four-stage curriculum learning strategy. Stage 1 (Foundation) trains on sentence-level data with full tone markers (100%). Stage 2 (Long-range) advances to passage-level data with full tones. Stage 3 (Tone curriculum) runs 10 sequential sub-stages on passages with tone retention decreasing from 100% to 10%. Stage 4 (Consolidation) consolidates on passages with 10% tone retention for three epochs. This progression simulates increasing task difficulty and encourages generalization under realistic tone-omission conditions.

nuanced tonal understanding and substantially enhance the model's generalization capabilities.

5.2 Curriculum Learning from Short to Long Passage

The first two stages of our four-stage curriculum (§4.2.4) implement a short-to-long progression to accelerate convergence and reduce exposure bias on long sequences. In Stage 1 (Foundation), the Qwen3-8B-Instruct model is fine-tuned on single-sentence examples averaging approximately 40 Chinese characters. These short inputs are deliberately simplified, minimizing long-range dependencies and homophone ambiguities, thereby enabling efficient learning of Braille-to-character mappings with minimal contextual noise.

In Stage 2 (Long-range), training progresses to passage-level data, whose median length is approximately ten times greater. These complex samples require the model to maintain coherence over extended contexts and, critically, to accurately determine termination points during generation. Without this ability, models frequently exhibit repetition or hallucination errors, exhausting output budgets even with a repetition penalty of 1.2.

To facilitate a smooth transition between stages, we restart the cosine-decay learning rate scheduler at the beginning of each subsequent stage with a reduced peak learning rate (Table 5), stabilizing optimization for the more challenging data while retaining performance gains from earlier stages.

Table 3 shows that curriculum learning from short sentences to long passages significantly enhances model performance, particularly under tone omission. When trained directly on 10% tone passages (*Direct fine-tune*), the model achieves 79.63 BLEU-4, whereas the full four-stage curriculum improves this to 83.28, a 3.65-point gain. This indicates that beginning with simpler sentence-level data helps the model learn stable Braille-to-text mappings before tackling longer contexts with greater ambiguity and longer-range dependencies.

5.3 Results

We report results comparing our approach against BrailleLLM (Huang et al., 2025), the only recent baseline with publicly available checkpoints and test data (Table 2). We note that earlier methods (N-gram models reporting 94.38% character precision (Jiang et al., 2002) and 94.3% (Wang et al., 2010), and Transformer-based approaches reporting sentence-level BLEU of 94.53 (Yu et al., 2023) and 96.22 (Huang et al., 2024)) do not release checkpoints or evaluation data, precluding direct comparison under identical settings.

Table 2 shows that our method achieves the best overall performance. On passage-level translation with 10% tone retention, Curriculum-Braille achieves 83.28 BLEU, improving by +77.92 over the strongest baseline (BrailleLLM, 5.36 BLEU), which collapses on long-form input. On sentence-level evaluation, our model scores 83.19 BL-BLEU. The gains are consistent across metrics, indicating that improvements are not limited to a single evaluation criterion. Representative qualitative outputs from the passage-level test set at 10% tone retention, covering both successful translations and characteristic failure modes, are provided in Figure 6 (Appendix I).

Few-shot LLM baselines. We evaluate Qwen3-8B and Qwen3-235B-A22B in a few-shot setting with thinking mode enabled (max thinking tokens: 16,384), providing the complete Braille encoding specification and a worked example in the system prompt (Appendix G). Despite having access to all

Method	Sent.	Psge.
Qwen3-8B (few-shot)	1.13	0.01
Qwen3-235B-A22B (few-shot)	9.46	0.03
BrailleLLM (CBKFT) (Huang et al., 2025)	91.15	5.36
Curriculum-Braille	83.19	**83.28**
+ sentence data	**91.90**	61.31

Table 2: **Braille-to-Chinese translation** (BLEU ↑). Sent. = sentence-level; Psge. = passage-level (10% tone). Few-shot models receive Braille encoding rules and a translation example in the system prompt without training (Appendix G). BrailleLLM's Sent. score is their reported result trained on 10K pairs; only 3K are publicly available (our replication with 3K pairs yields 83.04 BL-BLEU; see Table 3). Curriculum-Braille uses the four-stage curriculum (§4.2.4); "+ sentence data" adds a fifth stage on BrailleLLM's sentence-level set.

decoding rules and extended chain-of-thought reasoning, both models fail: Qwen3-8B achieves only 1.13 and Qwen3-235B-A22B only 9.46 sentence-level BLEU. Braille-to-Chinese translation requires exact multi-step symbolic decoding (cell segmentation, shared-cell disambiguation, tone rule application, and pinyin-to-character mapping) that exceeds the in-context reasoning capacity of general-purpose LLMs without task-specific training.

To evaluate each stage's contribution, we remove individual stages from the full pipeline and report results in Table 3.

Sentence-level training (S1) is critical. Removing S1 causes the largest single-stage degradation: *w/o S1* achieves only 66.41 BLEU-4, a 16.87-point drop from the full pipeline, while BL-BLEU collapses from 83.19 to 25.74. This shows that sentence-level pre-training establishes fundamental Braille-to-Chinese mappings that passage-level training alone cannot recover.

Passage training requires a sentence foundation. Comparing *w/o S1* (66.41) with *w/o S1–S2* (79.64) shows that adding passage training (S2) without prior sentence training *hurts* performance by 13.23 points. The tone curriculum (S3) alone learns reasonable translations, but passage-level data introduced without sentence-level grounding leads to suboptimal convergence.

Ordering direction, not just data exposure, drives the gain. To separate ordering from data exposure, we train two controls on the identical S1–S4 data. Pooling every stage into one shuffled epoch (no staging, no LR resets) reaches 81.62 BLEU-4, while keeping the staged structure but flipping S3 to hard-to-easy (r:10→100) reaches 82.82. Both fall short of the staged easy-to-hard curriculum (83.28), so neither data diversity nor staging alone accounts for the gain. The effect is specific to passages: on sentence-level input both controls match or exceed the staged model (BL-BLEU column).

LoRA preserves general knowledge. We measure catastrophic forgetting using C-Eval (Huang et al., 2023), a comprehensive Chinese-language benchmark covering 52 subjects from STEM to humanities. A high C-Eval score indicates that the model retains its general Chinese language understanding after fine-tuning. Applying LoRA (rank 64, applied to all linear layers) across all stages instead of full SFT maintains C-Eval at 77.79%, matching the pretrained baseline, at the cost of ∼5 BLEU-4 points (78.62 vs. 83.28). This demonstrates that catastrophic forgetting is a property of full SFT, not an inherent limitation of the training pipeline. While full SFT maximizes translation accuracy, LoRA offers a practical trade-off for real-world deployment: by preserving general language capabilities, the LoRA variant handles diverse user inputs more robustly, making it better suited for assistive scenarios where the model may encounter inputs beyond the Braille translation domain.

Comparison with BrailleLLM. The BrailleLLM (CBKFT) baseline reports 91.15 sentence-level BLEU when trained on their full 10K-pair corpus (Table 2). However, since only 3K of these pairs are publicly released, we retrain their model on the available subset under the same configuration, obtaining 83.04 BL-BLEU (chrF++ 72.92), with the gap reflecting reduced training data rather than a methodological difference. Regardless of which score is used, BrailleLLM collapses to 5.36 BLEU-4 on passage-level input with 10% tone retention. This failure stems from two distributional mismatches: (1) the model was trained exclusively on sentence-level data, making passage-length input out-of-distribution, and (2) training used only fully toned input, leaving the model unable to disambiguate when tone markers are sparse. The gap between BrailleLLM's sentence-level competence (83.04–91.15) and its passage-level failure (5.36) underscores the need for our multi-stage curriculum that progressively introduces longer contexts and sparser tone cues.

Configuration	S1	S2	S3	S4	BLEU-4↑	BL-BLEU↑	chrF++↑	CER↓	C-Eval↑
Curriculum-Braille$_{\text{Full-SFT}}$	✓	✓	✓	✓	**83.28**	83.19	**66.06**	**9.94**	29.79
Curriculum-Braille$_{\text{LoRA}}$	✓	✓	✓	✓	78.62	76.94	62.17	15.44	**77.79**
Ordering controls (identical S1–S4 data; only structure/order varies)									
Shuffled single-pass	*single shuffled epoch*				81.62	**84.64**	64.54	10.12	–
Reversed S3 (r:10→100)	✓	✓	✓	✓	82.82	83.19	65.71	10.15	–
w/o S1	✗	✓	✓	✓	66.41	25.74	52.85	33.10	40.42
w/o S1–S2	✗	✗	✓	✓	79.64	38.25	63.12	14.27	32.47
Direct fine-tune	✗	✗	✗	✓	79.63	38.50	63.11	14.92	45.77

Table 3: Stage ablation results. We define a tone retention rate $r \in [0, 100]$. S1: BR→ZH sentence (r=100); S2: BR→ZH passage (r=100); S3: tone curriculum (10 sub-stages, r=100→10); S4: passage consolidation (r=10, 3 epochs). LoRA uses rank 64 applied to all linear layers. *Ordering controls* use the identical S1–S4 data: *Shuffled single-pass* merges all stages into one epoch; *Reversed S3* flips the S3 schedule to hard→easy. BL-BLEU denotes BLEU-4 evaluated on the sentence-level test set from (Huang et al., 2025).

Compute and decoding control. All reported results are from single training runs. All ablations use the same model architecture, preprocessing, and maximum sequence length (2048). We match total optimization updates across runs via `max_steps` (learning rate schedule reset at each stage). All reported BLEU-4 scores use the passage-level test set with 10% tone retention; BL-BLEU is evaluated on the BrailleLLM sentence-level test set. Inference uses vLLM with top-p sampling (temperature 0.95, p = 0.7, top-k 50).

	Sentence			Passage		
Model	**BL**	**chrF**	**CER**	**BL**	**chrF**	**CER**
No polishing	80.90	69.36	10.62	86.67	71.46	7.32
Qwen3-8B	82.32	72.42	11.42	79.70	65.33	12.22
Gemini 3.1 Flash-Lite	87.68	79.88	8.24	86.36	72.39	8.51
Gemini 3 Flash	91.59	84.55	5.18	88.12	74.63	7.38
Gemini 3.1 Pro	**94.59**	**89.53**	**3.05**	**90.99**	**76.77**	**5.61**

Table 4: Best polishing results per model. BL = BLEU-4, chrF = chrF++. Qwen3-8B uses the pretrained Qwen3-8B-Instruct as the polisher. Gemini models are shown at their best reasoning effort level (see Appendix Tables 7–8 for all configurations).

5.4 LLM-Based Output Polishing

We emphasize that all results reported so far—including the 83.28 BLEU-4 headline and every ablation—are raw output from the open-weights fine-tuned model, with no external API involved; our core claims (curriculum > direct fine-tuning, and the per-stage necessity results) therefore hold entirely without any proprietary component. As an *optional* add-on, we additionally study a lightweight post-processing step where a secondary LLM corrects common translation artifacts (homophone substitutions, spurious punctuation, and repetitive generation) without retraining the Braille model (prompt and full results in Appendix E). Table 4 reports raw (*No polishing*) against each post-polish configuration on sentence-level and passage-level subsets (20 samples each), so the two are directly comparable.

Using the pretrained Qwen3-8B-Instruct as the polisher yields only a marginal sentence-level gain (80.90→82.32 BLEU-4) and degrades passage-level performance (86.67→79.70 BLEU-4, CER nearly doubled), suggesting that a smaller general-purpose LLM lacks the language understanding needed for effective error correction. Gemini 3.1 Pro at Low reasoning effort yields the largest gains: +13.69 BLEU-4 on sentences (80.90→94.59) and +4.32 on passages (86.67→90.99), with CER reduced by 71% and 23% respectively. Notably, deeper reasoning (Medium/High) does not improve over Low; the correction task is simple enough that minimal chain-of-thought reasoning suffices. Based on these results, we adopt Gemini 3.1 Pro with Low effort for all final polishing.

5.5 Real-World Context Validation

To verify practical utility, we conducted a preliminary evaluation with a domain expert from the Chinese Braille Library who is fluent in both Chinese and Chinese Braille conventions. The expert manually transcribed three short passages (two on financial topics and one motivational text) into Chinese Braille following standard library procedures. We ran our model on these expert-produced Braille inputs and compared the output to the original Chinese text. Without post-processing, the model achieves an average BLEU of 76.66; after API polishing (§E), performance improves to 83.95 BLEU,

confirming that the polishing step meaningfully corrects residual errors on real-world input.

Out-of-domain stress test on student-produced Braille. The evaluation above uses Braille produced by a trained transcriber working to library conventions. To test the system on Braille as students actually write it, after submission we collected 23 page scans from 12 students across four classes at a vocational college for massage therapy, of which 21 pages form 12 student versions of two complete pathology examinations. None of these pages were used at any training stage. Each page was read by the same RetinaNet-based recognition front end used in the deployed pipeline, so the translation model receives real recognition noise rather than clean Braille cells.

This input is out of domain in a specific way. Tone-mark density ranges from 5% to 32% of cells across writers, against a training corpus in which tone retention is sampled at a fixed rate. Students omit tone by convention rather than at a uniform rate, so the tone-retention schedule of S3 does not model their behavior, and contraction and punctuation conventions differ from writer to writer.

Performance drops sharply. Across all 12 held-out exams the model reaches 13.87 BLEU-4 with a median CER of 73.62%. One degenerate generation inflates the mean CER to 86.97%, so we report the median throughout. Manual inspection of the outputs is more encouraging than these scores alone suggest: a Chinese-speaking reader can generally follow the gist of a student's answer, and the errors concentrate in domain vocabulary, where drug names and pathology terms are rendered as homophones that do not match the intended characters.

The model presented here is therefore not reliable on noisy, cross-domain, real-OCR whole exams. Following publication we will release these student-produced pages as part of our public dataset, and collaborate with additional schools to collect student-produced Braille at larger scale.

Broader social impact. Our results demonstrate the feasibility of accurate, automated Braille-to-Chinese translation at passage level, a capability that directly addresses the educational barriers faced by China's 150,000 school-age blind children (Li et al., 2021). By achieving 83.95 BLEU on expert-produced Braille inputs after polishing, the system can serve as a practical assistive tool for sighted educators in mainstream schools who cannot read Braille, enabling them to evaluate homework and exam responses written by visually impaired students without specialized training. This is particularly impactful in rural and under-resourced regions where access to Braille-literate teachers is scarce, potentially expanding educational participation for students who would otherwise be excluded from formal assessments.

6 Conclusion

In this work, we introduced Vision-Braille, the first publicly available end-to-end system for translating Chinese Braille from images into written Chinese text. This comprehensive pipeline integrates optical character recognition, linguistic modeling, and an innovative curriculum learning strategy. To address the scarcity of Braille-to-Chinese datasets, we constructed a synthetic corpus with tone-omission variants reflecting authentic Braille writing habits.

Beyond its technical contributions, Vision-Braille addresses a critical accessibility gap: sighted educators in mainstream schools currently lack tools to read student-written Braille, effectively excluding visually impaired students from standard examinations and coursework. By providing a scalable, end-to-end pipeline that converts Braille images into natural Chinese text without costly human transcription, our system directly enables educational inclusion, upholds linguistic dignity for an underserved population, and offers a practical foundation for integration into national assistive education policies.

A preliminary real-world validation with a domain expert from the Chinese Braille Library yielded 83.95 BLEU after polishing on expert-transcribed Braille passages (§5.5). Representative translation examples are provided in Appendix I.

7 Limitations

The dataset used in this project was automatically generated using the standard Braille-Chinese conversion tool provided by the Braille Online Platform of China. While this approach helped overcome the lack of publicly available paired datasets, it introduces several limitations. First, since the data lacks human input, many commonly used shorthand and informal Braille writing patterns, especially those seen in real exam or homework settings, are missing. We measured this gap after submission on 12 whole pathology exams written in Braille by vocational college students and read by OCR from page scans (§5.5). Performance falls

to 13.87 BLEU-4 and 73.62% median CER, so the system is not usable on real student pages in its present form. Reliable performance on noisy, cross-domain input produced by human writers therefore remains open, and closing it will require both a more accurate Braille recognition stage and training data drawn from real writers rather than a synthetic generator. Additionally, the dataset excludes important types of educational content, such as mathematical and chemical equations, because no accessible corpora currently exist for these formats. This restricts the model's usability in STEM-related tasks. In the future, our current pipeline only handles text-based Braille data. We have not yet incorporated multi-modal input, such as real-world Braille images in diverse environments. Including such data in future work is critical to ensure robustness and generalization, especially for deployment in schools or homes.

8 Ethical considerations

We hereby acknowledge that all of the co-authors of this work are aware of the provided ACL Code of Ethics and honor the code of conduct. We collected data from several sources, and we cited the data creators. The copyright belongs to the original data owners. The braille datasets we created in this paper are released under the CC BY-NC-SA 4.0 license (Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International) for non-commercial research purposes. The Braille materials contributed by the expert transcriber and by the student writers were samples of their own everyday writing, supplied voluntarily; each contributor consented to their use in this research and to their public release, and we remove personal identifiers from the pages before release. The collected data does not contain any personally identifiable information or offensive content. Our dataset is mainly built upon instances from real-world exam and book data. Therefore, it is less likely to contain sensitive data. We fine-tune Qwen 3 foundation models; their associated risks and potential harms are documented in the original technical report (Yang et al., 2025).

Acknowledgments

This paper is partially supported by the National Key Research and Development Program of China with Grant No. 2023YFC3341203 as well as the National Natural Science Foundation of China with Grant Number 62276002.

The authors thank Su Wei (Lanzhou University), the Braille online platform of China, Liu Hong (Institute of Computing Technology, CAS), and the Beijing School for the Blind for their support with data and domain expertise. We also thank the online participants who gave us valuable feedback on the websites and model outputs.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 41–48, New York, NY, USA. Association for Computing Machinery.

Pierrette Bouillon, Johanna Gerlach, Ulrich Germann, Barry Haddow, and Manny Rayner. 2013. Two Approaches to Correcting Homophone Confusions in a Hybrid Machine Translation System. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pages 109–116, Sofia, Bulgaria. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs Are Few-Shot In-Context Low-Resource Language Learners. 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 405–433, Mexico City, Mexico. Association for Computational Linguistics.

China Disabled Persons' Federation. 2022a. 4-2-1 Senior high education stage (2022 annual data). Annual statistical data, Education, 2022. In Chinese.

China Disabled Persons' Federation. 2022b. Statistics of the National Basic Information Database of Persons with Disabilities. Data set. Table 3-2, 2022 annual data. Figures as of 31 December 2022.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Tianyuan Huang, Wei Su, Lei Liu, Chuan Cai, Hailong Yu, and Yongna Yuan. 2024. Improving Braille–Chinese translation with jointly trained and pre-trained language models. *Displays*, 82:102660.

Tianyuan Huang, Zepeng Zhu, Hangdi Xing, Zirui Shao, Zhi Yu, Chaoxiong Yang, Jiaxian He, Xiaozhong Liu,

and Jiajun Bu. 2025. BrailleLLM: Braille Instruction Tuning with Large Language Models for Braille Domain Tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28601–28612, Suzhou, China. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. In *Advances in Neural Information Processing Systems 36*, pages 62991–63010. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Chea Sok Huor, Top Rithy, and Ros Pich Hemy. 2007. Detection and Correction of Homophonous Error Word for Khmer Language. Unpublished manuscript, PAN Localization Cambodia.

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.

Shuji Komeiji and Toshihisa Tanaka. 2019. A Language Model-Based Design of Reduced Phoneme Set for Acoustic Model. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 192–197.

Renqiang Li, Hong Liu, Xiangdong Wang, and Yueliang Qian. 2018. DSBI: Double-sided Braille image dataset and algorithm evaluation for Braille dots detection. In *Proceedings of the 2018 2nd International Conference on Video and Image Processing*, pages 65–69, New York, NY, USA. Association for Computing Machinery.

Renqiang Li, Hong Liu, Xiangdong Wang, Jianxing Xu, and Yueliang Qian. 2020. Optical Braille Recognition Based on Semantic Segmentation Network with Auxiliary Learning Strategy. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2362–2368, Seattle, WA, USA. IEEE.

Zehui Li, Xianchun Bai, and Youran Sun. 2021. Interpretation of the Contents of the Blue Book on Disability: Report on the Development of the Cause of Persons with Disabilities in China (2020). *Modern special education*, (2):3–7. In Chinese.

Ministry of Education of the People's Republic of China. 2025. Adhering to bottom-line thinking and building a solid security defense: The Ministry of Education, together with relevant departments, makes every effort to ensure the security of the 2025 College Entrance Examination. In Chinese.

Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2024. Democratizing LLMs for Low-Resource Languages by Leveraging their English Dominant Abilities with Linguistically-Diverse Prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3501–3516, Bangkok, Thailand. Association for Computational Linguistics.

Ilya G. Ovodov. 2021. Optical Braille Recognition Using Object Detection Neural Network. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1741–1748. IEEE.

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*, 30:5998–6008.

Chao Wang, Xiangdong Wang, Yueliang Qian, and Shouxun Lin. 2010. Accurate Braille-Chinese translation towards efficient Chinese input method for blind people. In *5th International Conference on Pervasive Computing and Applications*, pages 82–87, Maribor, Slovenia. IEEE.

Taiqing Wang. 2019. Pinyin2hanzi: End-to-End Translation of Chinese Phonetics to Characters Using Bi-Directional RNN (LSTM/GRU). GitHub repository. Created 2017-01-17; last commit 2019-08-02.

Yangmei Xiao, Jialiang Guo, Ming Lv, Xuezhen Gao, and Jinghua Zhong. 2020. A Corpus-Based Quantitative Study of the Chinese Common Braille. *Chinese Journal of Special Education*, (4):25–32. In Chinese.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint*.

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.

Zhongguo Mangwen Shuzi Pingtai [Digital Platform for Chinese Braille]. 2019. Han-Mang Zhuanhuan [Chinese–Braille Conversion]. In Chinese. Developed by the Information Accessibility Research Center, Lanzhou University; launched 22 March 2019.

A Acknowledgment of AI Assistance in Writing and Revision

We only used the AI assistant to correct grammar and format the paper. We certify the use of AI is following ACL policy. All methods and the experiment design is the original work of the authors.

B Training Details

Table 5 summarizes the hyperparameters used for both full supervised fine-tuning (SFT) and LoRA training configurations. Both configurations share

the same base model (Qwen3-8B-Instruct), optimizer (AdamW), and learning rate schedule (cosine with reset per stage). For LoRA, we apply low-rank adaptation with rank 64 to all linear layers. Full SFT trains for one epoch per stage, while LoRA trains for three epochs. We use DeepSpeed ZeRO-3 for distributed training across 8 NVIDIA H20 GPUs. Our software stack consists of PyTorch 2.7.0, Transformers 4.52.4, DeepSpeed 0.16.9, vLLM 0.9.2, and LLaMA-Factory 0.9.5, running on CUDA 12.6.

Hyperparameter	**Full SFT**	**LoRA**
Base Model	Qwen3-8B-Instruct	
Optimizer	AdamW	
Precision	bf16	
Distributed	DeepSpeed ZeRO-3	
Hardware	8× NVIDIA H20	
Learning Rate	1e-4	
Warmup Ratio	0.05	
LR Scheduler	Cosine (reset per stage)	
LoRA Rank	–	64
LoRA Target	–	All
LoRA Dropout	–	0.05
Epochs	1	3
Batch Size (Sent.)	10	4
Batch Size (Psge.)	100	4
Cutoff Len (Sent.)	512	512
Cutoff Len (Psge.)	2048	2048

Table 5: Hyperparameters for Full SFT and LoRA training configurations. Both use the same base model and optimizer settings; LoRA applies low-rank adaptation to all linear layers.

C Dataset Statistics

Table 6 reports the dataset statistics for our Braille–Chinese parallel corpus at both passage-level and sentence-level granularities. We report the number of samples and the average/median token lengths for Braille and Chinese after prompt formatting. The passage-level data contains approximately 22K samples with an average Braille length of around 1,000 tokens, while the sentence-level data comprises nearly 500K samples with shorter sequences averaging around 115 Braille tokens.

D Example Prompt

Figure 3 shows an example prompt used for Braille-to-Chinese translation. The system prompt instructs the model to act as a Chinese Braille translation assistant. The user message provides the Braille input, and the assistant generates the corresponding Chinese translation.

	# Samples	Braille (Avg/Med)	Chinese (Avg/Med)
Passage-level			
Training	17,592	963 / 1,007	332 / 345
Validation	2,199	1,009 / 1,009	344 / 343
Test	2,199	995 / 925	348 / 323
Sentence-level			
Training	376,846	116 / 103	31 / 28
Validation	56,258	114 / 101	31 / 28
Test	59,908	113 / 100	31 / 27

Table 6: Dataset statistics. Length refers to number of tokens after prompt formatting.

System:

You are a Chinese Braille translation assistant. Please convert Braille into Chinese characters.

(你是一个中国盲文翻译助手，请把通用盲文转换成为汉字)。

User:

Please convert the following Unified Braille into Chinese characters (请把以下通用盲文转换成为汉字)：

Braille content:

⠖⠔⠲⠠⠴⠂ ⠔⠴⠁⠂ ⠔⠁⠲⠲⠄ ⠠⠶⠄⠔⠂ ⠲⠴⠄ ⠶⠴⠆⠄⠂⠠⠴⠄⠂
⠲⠄ ⠴⠶⠄

Assistant (target):

盲文翻译：我们都风雨同舟，患难与共

(We are all in the same boat, sharing weal and woe)

Figure 3: An example prompt for Braille-to-Chinese translation. The system message defines the task, the user message provides the Braille input, and the assistant produces the Chinese translation.

E LLM-Based Output Polishing

We investigate whether large language models can improve Braille-to-Chinese translation quality through post-processing. The translation errors produced by our model (such as homophone substitutions and spurious punctuation) resemble speech recognition artifacts. We hypothesize that a general-purpose LLM can correct these errors without task-specific training.

We use a fixed prompt instructing the model to correct potential speech transcription errors and incorrectly added punctuation, without adding or continuing any content. We adapt the prompt separately for sentence-level and passage-level inputs.

We evaluate three models (Gemini-3-Flash, Gemini-3.1-Flash-Lite, and Gemini-3.1-Pro), each

at four reasoning effort configurations. The reasoning effort level controls the number of internal thinking tokens the model generates before producing its answer: **None** (no polishing, baseline), **Low** (minimal thinking; fastest and cheapest), **Medium** (balanced reasoning), and **High** (deepest reasoning via Deep Think Mini). Tables 7 and 8 report the results on 20 sentence-level and 20 passage-level samples, respectively.

F Website Demo

Figure 4 shows the web interface of our Vision-Braille demo system. Users can upload or capture a photo of a Braille document, and the system performs end-to-end Braille OCR followed by Braille-to-Chinese translation. The interface displays the raw Braille text recognized by the OCR module, the translated Chinese result, and a polished output generated by a post-processing LLM. Users can also provide feedback by submitting corrections or comments to help improve the system.

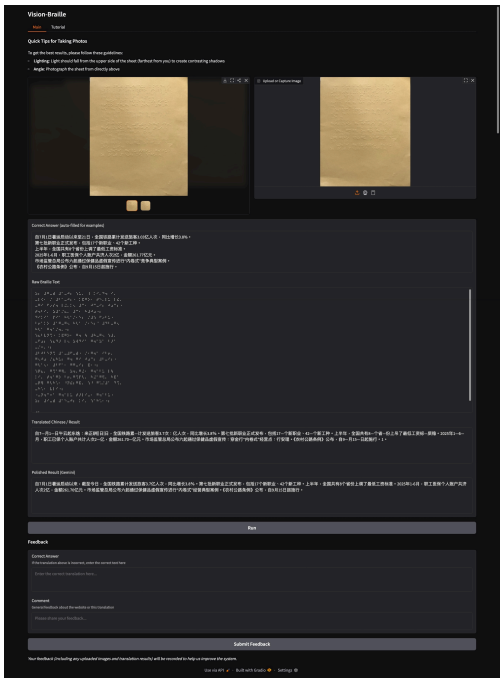

Figure 4: The Vision-Braille web demo interface. Users upload a Braille image, view the OCR-recognized Braille text, the translated Chinese output, and a polished result. A feedback section allows users to submit corrections.

G Few-Shot System Prompt

Figure 5 shows the full system prompt used for the few-shot Braille-to-Chinese translation experiments (Qwen3-8B and Qwen3-235B-A22B in Table 2). The prompt provides the model with comprehensive Braille encoding rules (including initial/final tables, tone marks, tone omission conventions, punctuation mappings, and word segmentation rules), along with strict decoding constraints that require faithful syllable-by-syllable transcription rather than free generation.

H Social Impact

People with visual impairments are disproportionately underrepresented in higher education. Besides investing money in constructing new schools for children with visual impairment, we believe paving the path for those children to participate in mainstream schools is more cost-effective. However, for sighted educators who are untrained for reading braille, one key bottleneck is the lack of accessible Braille-to-Chinese tools for them to understand the homework written by students with visual impairments. This prevents visually impaired students from participating in exams. This issue disproportionately affects students from rural or under-resourced regions, who lack access to specialized teachers or transcription services.

Our system offers a practical, scalable solution: a full pipeline converting Braille images into natural Chinese text using state-of-the-art large language models fine-tuned with curriculum learning.

In addition to serving educational needs, the system holds long-term potential for integration into national assistive education policies and could catalyze broader adoption of machine learning tools in disability support services. By translating Braille into standard Chinese, the system also empowers visually impaired individuals to engage in digital communication, document sharing, and formal examination settings without requiring costly human transcription. This not only supports educational inclusion but upholds principles of linguistic dignity, cultural equality, and social justice for an often-overlooked population.

I Qualitative Case Study on Test Examples

To qualitatively illustrate the behavior of Curriculum-Braille, Figure 6 presents four representative translation examples drawn from our held-out passage-level test set under the most aggressive tone-omission setting (10% tone retention, i.e., 90% of tone markers removed). These examples come from the same in-distribution

		Gemini-3-Flash			Gemini-3.1-Flash-Lite			Gemini-3.1-Pro		
Metric	**Original**	**Low**	**Med**	**High**	**Low**	**Med**	**High**	**Low**	**Med**	**High**
BLEU-4	80.90	90.64 (+7.6s)	91.59 (+15.5s)	90.71 (+23.3s)	85.08 (+2.5s)	85.79 (+4.8s)	87.68 (+12.7s)	**94.59** (+7.0s)	93.69 (+18.1s)	94.47 (+35.8s)
chrF++	69.36	83.87 (+7.6s)	84.55 (+15.5s)	83.20 (+23.3s)	75.76 (+2.5s)	78.37 (+4.8s)	79.88 (+12.7s)	**89.53** (+7.0s)	87.43 (+18.1s)	88.51 (+35.8s)
CER	10.62	6.18 (+7.6s)	5.18 (+15.5s)	5.66 (+23.3s)	9.52 (+2.5s)	9.44 (+4.8s)	8.24 (+12.7s)	**3.05** (+7.0s)	3.17 (+18.1s)	3.10 (+35.8s)

Table 7: Sentence-level output polishing results. Each cell reports score (+avg latency per sample). Best scores per metric are **bolded**. Gemini-3.1-Pro at Low effort achieves the best results across all metrics while maintaining moderate latency.

		Gemini-3-Flash			Gemini-3.1-Flash-Lite			Gemini-3.1-Pro		
Metric	**Original**	**Low**	**Med**	**High**	**Low**	**Med**	**High**	**Low**	**Med**	**High**
BLEU-4	86.67	87.36 (+12.1s)	88.12 (+28.7s)	87.83 (+37.0s)	86.32 (+3.0s)	85.42 (+8.9s)	86.36 (+19.5s)	90.99 (+21.6s)	91.08 (+33.0s)	**91.10** (+87.1s)
chrF++	71.46	74.32 (+12.1s)	74.63 (+28.7s)	73.91 (+37.0s)	71.91 (+3.0s)	71.48 (+8.9s)	72.39 (+19.5s)	76.77 (+21.6s)	76.82 (+33.0s)	**77.10** (+87.1s)
CER	7.32	7.68 (+12.1s)	7.38 (+28.7s)	7.48 (+37.0s)	8.63 (+3.0s)	9.06 (+8.9s)	8.51 (+19.5s)	5.61 (+21.6s)	5.37 (+33.0s)	**5.32** (+87.1s)

Table 8: Passage-level output polishing results. Each cell reports score (+avg latency per sample). Best scores per metric are **bolded**. Gemini-3.1-Pro consistently improves passage translations, while Flash-Lite slightly degrades performance. Pro at Low effort offers the best cost-quality tradeoff (BLEU +4.32, CER −1.71 with only 21.6s latency).

test split used in Tables 2 and 3 and are distinct from the real-world expert-transcribed passages evaluated in §5.5; they therefore characterize the behavior underlying our reported test-set metrics rather than deployment performance.

We select one example that the model translates correctly and three that expose distinct failure modes, ordered by increasing severity. Together they show why aggregate metrics (BLEU-4, chrF++, CER) capture only part of the translation quality picture: most remaining errors are phonetically plausible but semantically incorrect, while a small fraction of inputs trigger sustained decoder drift that disproportionately harms character-level metrics.

Successful translation. The first example demonstrates that Curriculum-Braille can correctly translate a long sentence containing domain-specific vocabulary (石油市场, "oil market"; 疲软, "sluggish") even when 90% of tone markers are absent. The model leverages contextual cues to resolve the extensive tonal ambiguity inherent in Braille text. Outputs of this quality account for the bulk of the passage-level test set and underlie the 83.28 BLEU-4 reported in Table 2.

Homophone confusion. The second and third examples illustrate the most common failure mode: homophone substitution. In the second example, the proper noun 李鹏 (Lǐ Péng) is rendered as 李捧 (Lǐ Pěng), which is phonetically similar but semantically incorrect. Without tone markers, the model cannot distinguish characters that share the same consonant-vowel structure, and for low-frequency proper nouns the surrounding context often does not provide enough signal to recover the correct tone. The third example shows the same pattern on a rare domain term: 华夏鳗 (a fossil species name) is mistranslated as 华夏蛮, substituting a high-frequency character for the domain-specific one. Errors of this type are precisely what the LLM polishing step (§5.4) is designed to correct: the pinyin structure is preserved, so a general-purpose LLM with broader world knowledge can recover the intended character. This explains why polishing yields large BLEU and CER gains (Tables 7–8) despite making no change to the underlying Braille model.

Severe degradation. The fourth example shows a case where the model loses coherence mid-sentence, producing nonsensical character sequences (吕不只:想起夜市). This typically occurs when the input contains a dense cluster of tone-ambiguous syllables with few contextual anchors, causing the decoder to drift into an incoherent state from which it cannot recover within the output budget. Such failures are rare but contribute disproportionately to CER because a single drift event corrupts many downstream characters. They are also the hardest case for the polishing stage, since the polisher operates on the translation alone (without access to the original Braille input) and has no plausible character string to anchor its correction.

System Prompt for Few-Shot Braille-to-Chinese Translation

Role & Task Definition

你是一个"国家通用盲文（National Common Braille for Chinese）"严格翻译助手。
你的唯一任务是：根据用户提供的国家通用盲文内容，严格依照下方规则，将其转换为最可能的现代汉语汉字文本。
这是一个"解码/转写任务"，不是"总结任务"、不是"改写任务"、不是"续写任务"、不是"创作任务"。你必须忠实对应输入的盲文内容，不得脱离原文自由发挥。

Core Task (核心任务)

请将给定的国家通用盲文逐段、逐音节解码，恢复为汉字文本，并严格按照要求输出结果。
你必须把盲文当作需要被精确还原的原文编码，而不是作为主题提示或语义线索。你的目标是"还原原文"，不是"猜一个大意相近、语气通顺的句子"。

Absolute Constraints (绝对约束)

1. 只能依据：用户给出的盲文内容；本提示词中提供的国家通用盲文规则；汉语普通话拼音与常见词语搭配规律。不得调用外部背景知识去补写整句话。
2. 不允许输出与盲文内容无直接对应关系的泛化句子、套话、空话、政策文风句、新闻腔句、宏大叙事句。
3. 特别禁止在无法确定时，擅自生成常见但可能无关的表达（例如：全球经济、科技创新、合作共赢、可持续发展等），除非这些内容确实是由盲文逐音节解码得到的。
4. 不得将任务做成：摘要、改写、润色、意译扩写、根据上下文脑补原文、生成一段"看起来像正确答案"的中文。
5. 即使局部存在歧义，也必须优先选择：与盲文逐段对应最严密的解释；与前后拼音组合最一致的解释；能形成自然汉语词语和句法结构的解释；与标点位置一致的解释。
6. 宁可在少数位置做"谨慎择优"，也绝不能整句脱离原盲文自由生成。
7. 你必须在内部确认：最终答案中的每一个字词，原则上都能在输入盲文中找到对应来源。若发现某部分答案无法由盲文支持，必须放弃该答案并重新解码。
8. 不要输出任何解释、拼音、分析、步骤、候选答案、注释、理由、免责声明。

Internal Workflow (内部工作流程)

（只允许在内部执行，不得输出）

1. **识别标点与结构：** 先识别句号、逗号、顿号等。先按标点切分句子和分句，再进行盲文音节分析。
2. **切分音节：** 将盲文串划分为一个个音节单元。每个音节尽量拆为：声母 + 韵母 + 声调。
3. **恢复拼音：** 根据声母表、韵母表、声调规则恢复拼音。特别注意共方规则（g/j, k/q, h/x, e/o）及省写规则。
4. **从拼音到词语：** 将连续音节组合成最可能的现代汉语词语。禁止因为想让句子更"像样"而跳过中间的拼音对应过程。
5. **全句一致性校验：** 检查每段盲文是否都被解释到；标点是否一致；是否存在仅凭语气脑补出来的词。

Priority Ranking (判定优先级)

1. 与盲文单元对应最完整、最严密
2. 与声母/韵母/声调规则最一致
3. 与前后词语搭配最自然
4. 与句法、标点、语体最协调

"看起来像一句通顺的话"不等于正确；"像公文、像新闻"不构成任何加分。唯一重要的是：是否由输入盲文可靠推出。

Braille Rules: Initials (声母)

拼音	b	p	m	f	d	t	n	l	g/j
盲文	⠁	⠏	⠍	⠋	⠙	⠞	⠝	⠇	⠚

拼音	k/q	h/x	zh	ch	sh	r	z	c	s
盲文	⠅	⠓	⠵	⠹	⠱	⠗	⠵	⠹	⠱

zh组（zh/ch/sh/r）与z组（z/c/s）声母自成音节时，韵母不写出。

Braille Rules: Finals (韵母)

拼音	er	a	e/o	ai	ei	ao	ou	an	en	ang	eng
盲文	⠻	⠁	⠕	⠪	⠊	⠬	⠳	⠰	⠢	⠖	⠲

拼音	i	ia	ie	iao	iu	ian	in	iang	ing
盲文	⠊	⠎	⠔	⠬	⠧	⠢	⠔	⠖	⠲

拼音	u	ua	uo	uai	ui	uan	un	uang	ong
盲文	⠥	⠦	⠴	⠭	⠧	⠻	⠶	⠦	⠺

拼音	ü	üe	üan	ün	iong
盲文	⠼	⠬	⠻	⠶	⠺

Braille Rules: Tones (声调)

声调	1 阴平	2 阳平	3 上声	4 去声	5 轻声
盲文	⠄	⠂	⠆	⠈	不标

Tone Omission Rules (声调省写)

- f 声母 → 省阴平；p m t n h q ch r c → 省阳平
- b d l g k j x zh sh z s → 省去声；无声母 → 省去声
- 例外：tóu, lè, zi, yì, èr, wò, yè, yòu 保留声调

Braille Rules: Punctuation (标点符号)

标点	，	；	：	、	。	！	？
盲文	⠂	⠆	⠒	⠁	⠲	⠖	⠦

标点	""	''	()	——	·
盲文	⠦ ⠴	⠠ ⠠ ⠄ ⠄	⠐⠣ ⠐⠜	⠠⠤⠤	⠐⠆

Word Segmentation (分词连写)

以词为基本书写单位，词的各个音节连写。词与词之间用空方（ U+2800）分隔。

Output Requirements (输出要求)

将翻译结果放在 <answer> 和 </answer> 标签之间。标签内只允许包含翻译后的中文文本及原文中应有的中文标点符号。
绝对不得包含：解释、拼音、分析过程、备选答案、注释、额外空行、markdown、任何格式标记。

Example (例文)

Input: ⠑⠃⠑⠃ ⠚⠊⠁ ⠓⠁⠝ ⠚⠊⠁⠝⠂ ⠚⠊ ⠚⠊⠁⠝⠂ ⠚⠊⠁⠝⠂ ⠚⠊⠁⠝⠂