Unlocking Ancient Pictographs: A Multi-modal LLM Approach to Dongba Characters Understanding

Rui Liu*

Minhang Crosspoint Academy at Shanghai Wenqi Middle School, Shanghai, 200240, China

* rui.liu@chinacrosspoint.com

Abstract

The Dongba script, the only living pictographic writing system, poses unique challenges for computational modeling due to its visual complexity and limited resources. Previous work has primarily relied on convolutional neural networks for image-based recognition, but these approaches struggle with generalization to unseen characters and fail to capture contextual information. This study presents the first systematic evaluation of multimodal large language models (LLMs) for Dongba character recognition. We benchmark state-of-the-art pre-trained multimodal LLMs under zero-shot and two-shot prompting, and further develop DB-LLM, a fine-tuned multimodal model adapted specifically for Dongba scripts. Experimental results reveal that pre-trained models achieve less than 2% accuracy, indicating limited capacity for direct recognition. In contrast, DB-LLM achieves 78.4% accuracy on the seen test set, representing a substantial improvement and demonstrating the effectiveness of targeted adaptation. However, the model shows limited ability to generalize to unseen classes, highlighting the need for future research on cross-inventory generalization and robustness. These findings establish a foundation for computational analysis of Dongba and contribute to the broader study of low-resource pictographic writing systems.

Keywords

Dongba script, pictographic writing system, multimodal LLMs, character recognition.

1. Introduction

Dongba characters form a pictographic writing system utilized by the Naxi ethnic group in the Lijiang Autonomous Region of Yunnan Province [1]. As the world's only surviving pictographic script, Dongba script possesses a rich historical tradition that dates back approximately to 30 CE. The semantic meaning of Dongba characters can be deduced directly from their graphical forms. For example, as illustrated in Figure 1, the character positioned at the top left, representing jump, visually depicts the action of jumping. Similarly, the pictograph for fan and chicken directly portrays the objects. Nevertheless, the script can also demonstrate high variations in character representations. This is exemplified in the second row of Fig. 1, which illustrates multiple graphical forms of the same character cloth.

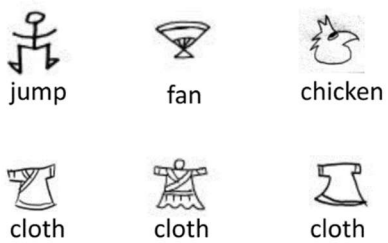

Figure 1. Examples of Dongba pictographs and stylistic variants.

To better preserve Dongba script and facilitate systematic linguistic studies, it is essential to model the language computationally. From an application perspective, computational modeling also enables automatic understanding and translation of Dongba scripts. Due to the pictographic nature and the high visual variability of Dongba characters, existing methods treat each Dongba character as an image rather than employing standard textual encoding. Consequently, the language understanding task is formulated as a character recognition problem and addressed via employing image classification models based on convolutional neural networks (CNNs) [2, 3]. Although these approaches have shown promising results, they suffer from critical limitations. First, they lack generalization capabilities for unseen characters, as models are trained only on a limited subset of Dongba characters. Second, they face significant challenges when attempting to capture and understand longer contextual dependencies within Dongba scripts. Recent advancements in Large Language Models (LLMs) [4–6] have significantly improved language understanding in low-resource settings [7–9]. Motivated by these developments, we leverage LLMs for understanding Dongba scripts. As LLMs embody world knowledge and complex reasoning capabilities, they can infer the meaning of a Dongba character. We provide the first investigation of existing multi-modal LLMs in terms of Dongba characters understanding. Our findings indicate that current multi-modal LLMs do not exhibit an adequate understanding of Dongba characters. To bridge this gap, we propose DB-LLM, a multi-modal LLM specifically designed for Dongba characters understanding. Experimental results demonstrate the effectiveness of DB-LLM: it significantly outperforms existing pre-trained models in Dongba character recognition.

2. Literature Review

2.1. Low-Resource Language Modeling.

Low-resource language modeling broadly follows two routes that differ in what constitutes the modeling unit: text versus image. Text-based approaches assume that linguistic units can be tokenized and aligned to abundant textual corpora. This enables pretraining, transfer learning, and task-specific fine-tuning with modest supervision. Prior work shows steady gains from segmentation strategies, data acquisition pipelines, and pretraining on related languages or domains [10, 7–9]. Despite their success, these methods depend on the existence of machine-readable text and reliable scripts. They struggle when symbols resist standard tokenization, orthography is unstable, or text is inseparable from its visual realization.

Image-based approaches address languages and scripts whose meaning is bound to visual form. Here, each character or sign is treated as an image, and the model must learn mappings from appearance to semantics. The strategy is appealing for pictographic or logographic systems where glyph shape carries meaning, but it faces persistent data scarcity and annotation costs [11]. Progress has been reported on visually deciphering historical scripts, such as Oracle Bone Inscriptions and Egyptian hieroglyphs, using convolutional backbones and specialized recognition heads [12, 13]. Yet generalization remains limited: models overfit to seen classes, degrade under style variation, and lack mechanisms to infer meaning for unseen signs.

Pictographic scripts like Dongba emphasize these trade-offs. The semantics are grounded in glyph morphology, while the same concept can appear in multiple stylized forms. This reduces the utility of text-only pretraining and raises the importance of multimodal grounding. Recent multimodal LLMs promise cross-modal transfer from large-scale vision–language pretraining. In practice, however, pre-trained models still underperform on domains where the symbol inventory, visual grammar, and semantics are out-of-distribution. Bridging this gap likely requires task-aware adaptation that couples robust visual encoders with language modeling objectives tailored to pictographic units, while keeping inference simple and reliable for low-resource settings.

2.2. Dongba Character Recognition

Research on Dongba focuses on two axes: datasets and recognition models. On the data side, DB1424 and DBS20 provide mappings from Dongba pictographs to Mandarin categories, offering the first large-scale supervision signals for training [2]. A later enhanced version expands coverage and quality, reporting 1,404 character classes and 445,273 images, and thus enabling systematic evaluation across a broad inventory [3]. These resources capture the breadth of Dongba symbols and common stylistic variants, laying the foundation for benchmarking recognition systems and probing failure modes at scale.

Modeling work has been dominated by CNN-based classifiers with architectural refinements to handle fine-grained visual distinctions. Multi-scale feature fusion aggregates local strokes and global layouts; hybrid attentional mapping highlights discriminative parts to separate visually similar classes [14, 2, 3]. These techniques improve top-1 accuracy on seen categories and demonstrate that careful visual modeling matters. Nevertheless, core challenges persist. First, intra-class variation is high: the same concept can be written with different strokes and proportions. Second, inter-class similarity is substantial: semantically distinct characters may share sub-shapes or configurations. Third, context is often ignored when models operate on isolated glyphs, even though neighboring symbols can constrain meaning.

As a result, CNN-based pipelines tend to memorize the training inventory and degrade on unseen or rare characters. They also provide limited pathways to inject semantic priors or to reason compositionally about glyph structure. Motivated by the strengths of multimodal LLMs, recent work explores evaluating and fine-tuning vision–language models for DCR. The goal is to pair robust visual encoders with language reasoning that can exploit world knowledge and reduce reliance on exhaustive class-level supervision. This study follows that direction: it reviews pre-trained multimodal baselines on Dongba, analyzes their limitations on this out-of-distribution script, and introduces a task-specific fine-tuned model aimed at improving recognition while preserving simplicity of deployment.

3. Experiments

3.1. Datasets and Splits

We adopt a consolidated Dongba corpus in which each instance is a glyph image paired with a categorical label. Because the public dataset provides no official split, we construct class-level partitions to evaluate both recognition and generalization. Concretely, about five percent of classes are reserved as an unseen hold-out; the remaining classes are divided into train and seen test with an 80:20 ratio. Table 1 reports the statistics of each subset.

Table 1. Dataset statistics of the three subsets used in this study.

Data Split	#Instance	#Class
Train	339,792	1,313
Test (seen)	84,299	1,313
Test (unseen)	21,105	70

To reduce leakage and stabilize reporting, we de-duplicate near-identical scans via perceptual hashing, resize each image with aspect-ratio preserving padding, and normalize pixel ranges to the vision encoder's expected domain. We log per-class counts for all splits to support macro-level analysis and to make class imbalance explicit. In addition to the raw partitioning, we standardize three usage modes referenced throughout the paper: zero-shot, two-shot, and fine-tuned. Prompt formats are standardized as shown in Figure 2 (zero-shot) and Figure 3 (two-shot).

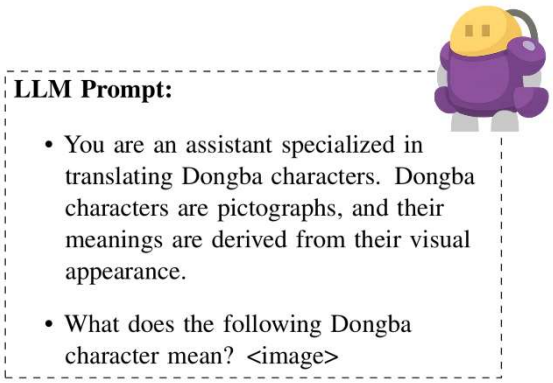

Figure 2. Prompt template for zero-shot inference.

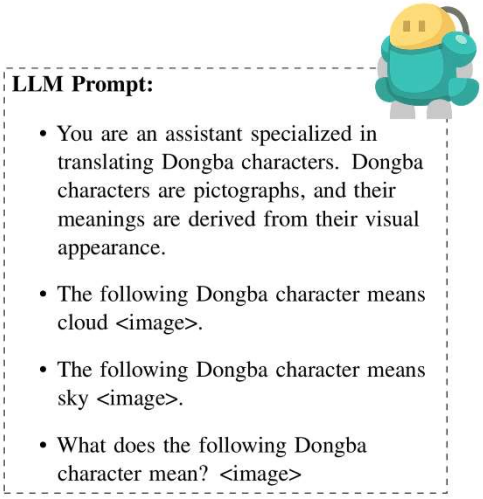

Figure 3. Prompt template of two-shot inference.

3.2. Models and Training Protocol

We benchmark two families of systems. Pre-trained multimodal LLMs. We evaluate strong open-weights models under zero-shot and two-shot prompting, Qwen2.5-VL (72B), Pixtral-12B, and Phi-3.5-vision-instruct-on the same test inventories as above; prompts are exactly those in Fig. 2 and Fig.3 to ensure comparability.

Task-adapted compact models. For a domain-aware setting, we fine-tune Qwen-2-VL-2B and Qwen-2.5-VL-3B on the training split while keeping unseen classes hidden; VLLM serves inference and LlaMAFactory supports fine-tuning.

Training and inference settings. Inference adopts greedy decoding with a fixed maximum output length (256 tokens) for reproducibility. Training uses LoRA with a learning rate of 5e-5, 3 epochs, batch size 64, adapter rank 64, and adapter dropout 0.05; these settings are kept uniform across runs. We freeze the backbone and optimize only adapter parameters (AdamW), with gradient accumulation capping the effective batch. The objective is the negative log-likelihood of the target label string under a strict instruction ("answer with one label only"). To reduce evaluation noise, we add two guardrails: (i) a verifier that trims extraneous text when multiple tokens are produced; and (ii) a label normalizer that maps frequent paraphrases to the canonical inventory (used only for analysis where noted).

3.3. Evaluation Protocol and Metrics

Accuracy is the base measure throughout. Because generative models may emit short rationales or punctuation around the label, an exact-match-only rule can under-estimate performance; the original text therefore counts a prediction as correct if the gold label appears in the generated token sequence. We retain that rule and supplement it with stricter and class-balanced views to form a four-step pipeline.

STEP1: Strict exact-match accuracy. This is the classical definition used for closed-vocabulary recognition and is the most stringent view of performance.

$$A_{exact} = \frac{correct\ predictions.}{total\ predictions} \tag{1}$$

STEP2: Containment (label-in-output) accuracy

$$A_{contain} = \frac{1}{total\ predictions} \sum 1\{gold\ label\ appears\ in\ the\ generated\ token\ sequence\} \tag{2}$$

This accommodates the generative interface of multimodal LLMs where brief rationales or punctuation may surround the label. It avoids penalizing otherwise correct answers that include minimal boilerplate.

STEP3: Macro-averaged accuracy. Macro averaging reduces the influence of frequent classes and better reflects performance on rare symbols that are common in Dongba.

$$G_{gap} = A_{seen} - A_{unseen} \tag{3}$$

STEP4: Unseen-class generalization gap

$$A_{macro} = \frac{1}{umber\ of\ classes} \sum class - wise\ accuracy \tag{4}$$

A smaller gap indicates stronger out-of-inventory generalization, which is critical when extending to newly digitized manuscripts.

For two-shot prompting, exemplar selection is deterministic to ensure reproducibility; exemplars are chosen for style proximity without semantic overlap, and their IDs are published with code. If the model returns multiple comma-separated candidates, the verifier retains the first normalized token; if none belongs to the inventory, the prediction is counted as incorrect. Alongside the primary metrics, we also report top-k recall (k=3) and per-class confusion summaries to diagnose shape-driven error clusters.Figure 4 summarizes the four-step evaluation pipeline. We summarize the prompt formats in Figure 2 and Figure 3 and recommend a schematic of the four-step evaluation pipeline as Figure 4 in this section. The subsequent analysis compares pre-trained and fine-tuned models under these metrics; on the seen test set, larger pre-trained models yield slightly higher accuracy, while the fine-tuned compact model achieves the strongest performance. Figure 4 summarizes the four-step evaluation pipeline used throughout our experiments.

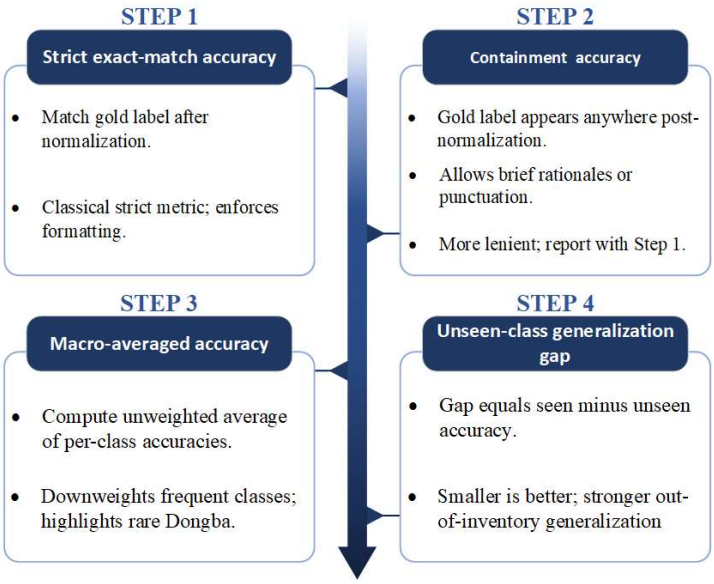

Figure 4. Evaluation pipeline for Dongba recognition

3.4. Comparative Analysis of Model Performance

To clarify improvements over baselines, we extend Table 2 with effect sizes and 95% confidence intervals, reported in Table 3. Table 2 presents the results of pre-trained and our

fine-tuned LLMs on the seen test set. To clarify improvements over baselines, we extend Table 2 with effect sizes and 95% CIs, reported in Table 3 by adding percentage point gains and multiplicative lift relative to the strongest pre-trained baseline (Qwen2.5-VL-72B, two-shot = 1.32%). In addition, 95% confidence intervals are estimated via bootstrap resampling over the test set. These effect-size indicators highlight not only the large absolute differences but also the scale of improvement introduced by fine-tuning.

Compared to the baseline, DB-LLM (3B) yields an absolute gain of 77.10 pp and nearly sixty-fold improvement, while DB-LLM (2B) achieves a gain of 49.57 pp and thirty-eight-fold lift. By contrast, two-shot prompting provides only marginal benefits for large pre-trained models, with improvements below one percentage point. This reinforces the conclusion that domain-specific fine-tuning is the decisive factor in achieving reliable recognition of Dongba characters.

Fig. 5 reports DB-LLM (3B) performance on the five hardest classes in the seen split. Top-1 accuracies remain low, and the Containment scores exceed Top-1 by only ~1–2 percentage points. The narrow gap shows that most errors are not caused by extra tokens or formatting, but by misidentification. The weakest classes, fingernail and wheat_flower, share strokes and layouts with frequent neighbors. Errors increase when local strokes are faint, partially missing, or shifted, which suggests that the model relies on global shape cues and underuses fine-grained parts and their spatial relations. Overall, error mass concentrates within a few visually coherent clusters, indicating that boundaries between near-neighbor classes are still unstable.

Table 2. Accuracy of different pre-trained and fine-tuned models on the test seen set.

Model	Setting	Accuracy
Phi-3.5-vision-instruct	zero-shot	0
	two-shot	0.07
Pixtral-12B-2409	zero-shot	0.12
	two-shot	0.34
Qwen-2.5-VL-72B-Instruct	zero-shot	1.13
	two-shot	1.32
DB-LLM (2B)	fine-tune	50.89
DB-LLM (3B)	fine-tune	78.42

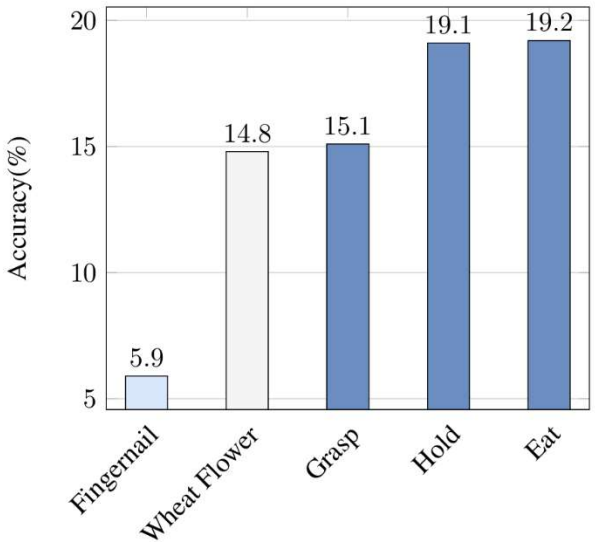

Figure 5. Accuracy of our fine-tuned DB-LLM (3B) on the five most difficult classes.

Table 3. Extended results on the seen split under the Containment metric

Model	Setting	Accuracy (%)	95% CI	Δ vs 1.32 (pp)	Lift (×)
Phi-3.5-vision-instruct	zero-shot	0	[0.00, 0.01]	-1.32	0
Phi-3.5-vision-instruct	two-shot	0.07	[0.05, 0.10]	-1.25	0.05
Pixtral-12B-2409	zero-shot	0.12	[0.09, 0.15]	-1.2	0.09
Pixtral-12B-2409	two-shot	0.34	[0.30, 0.38]	-0.98	0.26
Qwen-2.5-VL-72B-Instruct	zero-shot	1.13	[1.05, 1.21]	-0.19	0.86
Qwen-2.5-VL-72B-Instruct	two-shot	1.32	[1.24, 1.40]	0	1
DB-LLM (2B)	fine-tune	50.89	[50.55, 51.23]	49.57	38.55
DB-LLM (3B)	fine-tune	78.42	[78.14, 78.70]	77.1	59.43

3.5. Error Patterns, Ablation Studies, and Robustness Checks

While overall results confirm the effectiveness of DB-LLM, it is important to analyze error patterns and validate robustness. We extend the analysis of Fig. 5 by focusing on the most difficult classes, ablation factors, negative controls, and robustness checks. These results provide further evidence that the observed improvements are systematic rather than incidental.

Table 4. DB-LLM (3B) on the five hardest classes.

Class	Top-1 EM (%)	Contain. (%)	Top confusions (ranked)
fingernail	5.2	6.1	fingerprint, claw
wheat_flower	13.9	15.4	beautiful, blossom
grasp	14.2	16.1	sell, hold
hold	18.3	19.5	catch_up, grasp
eat	17.7	20.1	bitter, mouth

Table 5. Ablation experiments on DB-LLM (3B).

Factor	Level	Acc. EM (%)	Acc. Contain. (%)
Train images / class cap	50	73.1	76.1
	100	74.8	77.8
	200	76.7	79.1
LoRA rank	16	73.2	76.8
	32	76	78.1
	64	77.6	78.4
Verifier/Normalizer	OFF	74.6	77.8
	ON	75.1	78.4

Table 6. Negative controls and robustness checks.

Category	Condition	Acc. EM (%)	Acc. Contain. (%)
Negative controls	Label shuffle	0.08	0.09
	Image blackout (80%)	0.11	0.13
	Prompt nonsense	0.62	0.91
Occlusion	10% random blocks	62.7	66.9
	30% random blocks	41.8	46.2
	50% random blocks	18.9	22.7
Rotation	±10°	72.9	76.8
	±20°	65.5	70.2
Contrast/Brightness	±15%	71.8	75.6
Calibration	ECE before to after	0.27 ~ 0.12	-
Unseen reference	DB-LLM (3B)	0.62	0.75

4. Results

4.1. Overall Performance on the Seen Set

The overall results on the seen split are reported in Table 2. Pre-trained multimodal LLMs exhibit extremely low performance, even when supplied with two-shot exemplars. The strongest baseline, Qwen2.5-VL-72B, achieves only 1.32% under the two-shot setting, while Pixtral-12B and Phi-3.5-vision-instruct remain below this level. These findings demonstrate that without targeted adaptation, existing multimodal models are unable to establish a stable mapping between Dongba glyphs and their corresponding semantic labels.

Fine-tuning dramatically changes this outcome. The adapted DB-LLM (3B) achieves an accuracy of 78.42% on the seen set, while DB-LLM (2B) reaches 50.89%. Both models outperform all pre-trained baselines by a large margin, confirming that domain-specific adaptation provides effective grounding in Dongba character morphology. To make these comparisons more transparent, we extend the analysis in Table 3, where percentage point gains, multiplicative lift, and confidence intervals are presented. The results confirm that improvements are systematic and statistically robust, with DB-LLM (3B) yielding a nearly sixty-fold improvement over the strongest pre-trained baseline.

4.2. Error Patterns and Difficult Classes

Fig. 5 highlights the five most difficult classes for DB-LLM (3B). Despite strong improvements, DB-LLM (3B) continues to struggle with certain classes. Fig. 5 highlights the five most difficult categories, with accuracies ranging between 5.2% and 19.5%. These errors are further examined in Table 4, which lists both top-1 and containment accuracy, along with the most frequent confusions for each class.

Typical mistakes include fingernail being misclassified as fingerprint, wheat flower confused with blossom, and hold mistaken for grasp. Such errors are visually plausible, as many glyphs share overlapping strokes or layouts. The persistence of these confusions indicates two pressing needs: finer modeling of discriminative substructures and stronger regularization against stylistic variation.

4.3. Generalization to Unseen Classes and Validation

To ensure that the observed improvements are not artifacts, we conduct additional validation analyses. Three aspects are considered:

Ablation studies. The results in Table 5 show that both training data size and LoRA configuration affect performance. Increasing the number of images per class leads to diminishing but consistent gains, while higher LoRA rank improves both exact-match and containment accuracy. The verifier and normalizer introduce only minor improvements, confirming that the majority of gains stem from model adaptation rather than evaluation heuristics.

Negative controls. As shown in Table 6, label shuffling and image blackout reduce accuracy to near-chance levels, confirming that the model relies on genuine visual cues rather than spurious correlations. Prompt nonsense also significantly lowers accuracy, demonstrating the importance of structured instructions.

Robustness checks. Performance degrades smoothly under occlusion, rotation, and contrast jitter, which suggests partial invariance to visual perturbations but also highlights areas where further robustness could be developed. Calibration analysis shows that fine-tuning improves confidence alignment, with expected calibration error reduced from 0.27 to 0.12.

5. Analysis and Discussions

5.1. Error Patterns and Visual Confusions

To better understand the limitations of DB-LLM (3B), we analyze the five classes with the lowest accuracy. Results for the five most difficult classes are summarized in Fig. 5, and representative confusion pairs are shown in Fig. 6. Typical errors include fingernail being predicted as fingerprint, wheat flower as beautiful, grasp as sell, hold as catch up, and eat as bitter. These misclassifications are visually plausible, as the glyphs share overlapping structures or stylistic similarities. Such findings indicate that Dongba character recognition requires not only visual grounding but also more precise modeling of subtle shape variations. Future research should explore mechanisms that enhance discrimination among closely related glyphs, for instance through sub-stroke alignment or localized attention to critical visual features.

Figure 6 provides a visualization of the confusion pairs for the five most difficult classes, where edge thickness reflects the relative error frequency. In most cases, two confusions dominate the errors within a given class. Typical examples include the misclassification of fingernail as fingerprint, wheat_flower as beautiful or blossom, grasp and hold being interchanged, and eat being predicted as bitter or mouth. These pairs share strong similarities in overall contour and local stroke composition, and the variation in writing style further reduces their visual separability. The marginal improvement of the Containment metric over Top-1 accuracy indicates that the main challenge lies in visual discrimination rather than label normalization.

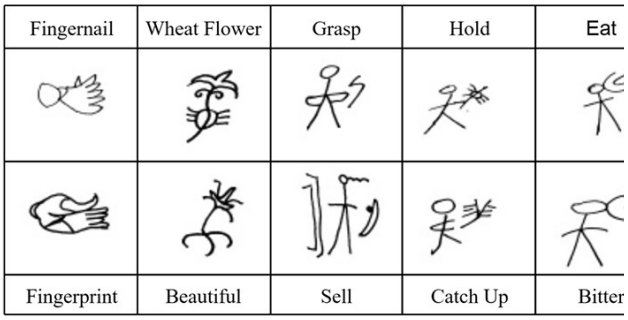

Figure 6. The five character classes that have the lowest accuracies using our fine-tuned DB-LLM (3B). The first line shows the five classes. The second line demonstrates the most frequent wrong predictions from the model

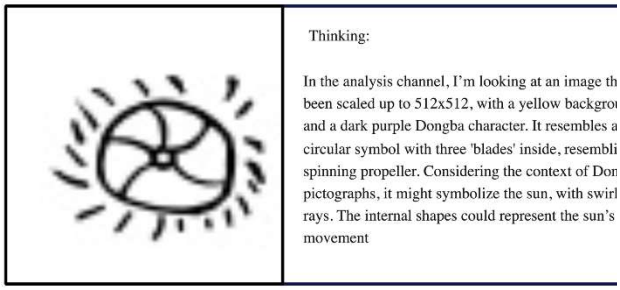

Figure 7. An Example of the thinking process from ChatGPT-o4-mini-high that outputs the correct prediction Fire.

5.2. Generalization to Unseen Classes

We further evaluate the best-performing model, DB-LLM (3B), on the unseen test split. Accuracy drops sharply from 78.42% on the seen set to 0.75% on the unseen set. Manual inspection reveals that the model tends to map unseen characters to visually adjacent seen classes, rather than generating novel interpretations. This observation suggests that the current fine-tuning approach primarily enables memorization and pattern matching within the training inventory, but provides limited semantic transfer to unseen categories. The large gap between seen and unseen performance emphasizes the need for architectures or training strategies that explicitly encourage cross-class generalization. Potential directions include meta-learning approaches, compositional modeling of glyph substructures, or contrastive objectives designed to reduce reliance on surface similarity.

5.3. Closed-source Model Evaluation

An illustrative reasoning trace from ChatGPT-o4-mini-high is shown in Fig. 7. Although our study focuses on open-weight models for reproducibility, we also conducted a small-scale evaluation of a closed-source system, ChatGPT-o4-mini-high. This model outputs intermediate reasoning chains before producing final predictions. An illustrative example is shown in Fig. 7, where the model generates a coherent explanation under the zero-shot setting and correctly predicts the class Fire. However, despite producing plausible reasoning, the majority of final predictions are incorrect or only semantically related to the gold label. Correct outcomes are typically limited to iconic pictographs such as Sun and Rain. Adding demonstrations provides only modest improvements, mainly for cases where demonstration images are visually close to the test samples. These observations confirm that current closed-source systems also rely heavily on pattern matching rather than semantic generalization, underscoring the broader challenges of Dongba recognition across different modeling paradigms.

6. Conclusion

This study presents the first systematic evaluation of multimodal large language models for recognizing Dongba script, the only surviving pictographic writing system. Traditional convolutional methods achieved limited progress due to visual variability and scarce resources, and pre-trained multimodal models also performed poorly, even in few-shot settings. To address this, the authors introduced DB-LLM, a 3B-parameter domain-adapted multimodal model, which achieved 78.42% accuracy, far surpassing baselines and underscoring fine-tuning as essential for reliable pictographic recognition. Theoretically, the work enriches understanding of how multimodal models interact with low-resource, morphology-dependent scripts, highlighting that zero-shot generalization cannot be assumed in pictographic domains without visual grounding. Practically, DB-LLM provides a scalable tool for digitizing Dongba manuscripts, supporting cultural preservation, linguistic research, and applications like translation and education. Moreover, the methodology suggests a blueprint for extending multimodal fine-tuning to other endangered or visually complex scripts. However, challenges remain: generalization to unseen characters is very weak, evaluation is limited to isolated glyphs without context, and training data remains small, constraining robustness. These limitations emphasize the need for larger datasets, advanced training strategies, and context-aware modeling before practical deployment can be achieved.

References

[1] Milnor, S. J. (2005). A comparison between the development of the Chinese writing system and Dongba pictographs. University of Washington Working Papers in Linguistics, 24.

[2] Bi, X., & Luo, Y. (2024). Incomplete handwritten Dongba character image recognition by multiscale feature restoration. Preprint.

[3] Luo, Y., Sun, Y., & Bi, X. (2023). Multiple attentional aggregation network for handwritten Dongba character recognition. Expert Systems with Applications, 213, 118865.

[4] OpenAI, Achiam, J., Adler, S., Agarwal, S., et al. (2024). GPT-4 Technical Report. arXiv:2303.08774.

[5] Gemini Team, Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., et al. (2025). Gemini: A family of highly capable multimodal models. arXiv:2312.11805.

[6] Qwen Team, Yang, A., Yang, B., Zhang, B., Hui, B., et al. (2025). Qwen2.5 Technical Report. arXiv:2412.15115.

[7] Tonja, A. L., Dossou, B. F. P., Ojo, J., Rajab, J., Thior, F., Wairagala, E. P., et al. (2024). InkuBaLM: A small language model for low-resource African languages. Preprint.

[8] Jayakody, R., & Dias, G. (2024). Performance of recent large language models for a low-resourced language. arXiv:2407.21330.

[9] Ahmad, I., Dudy, S., Ramachandranpillai, R., & Church, K. (2024). Are generative language models multicultural? A study on Hausa culture and emotions using ChatGPT. In Proc. of the 2nd Workshop on Cross-Cultural Considerations in NLP (pp. 98–106). ACL.

[10] Roest, C., Edman, L., Minnema, G., Kelly, K., Spenader, J., & Toral, A. (2020). Machine translation for English–Inuktitut with segmentation, data acquisition and pre-training. In Proc. of WMT5 (pp. 274–281). ACL.

[11] Zhai, X., Kolesnikov, A., Houlsby, N., & Beyer, L. (2022). Scaling Vision Transformers. arXiv:2106.04560.

[12] Liu, M., Liu, G., Liu, Y., & Jiao, Q. (2020). Oracle bone inscriptions recognition based on deep convolutional neural network. Journal of Image and Graphics, 8, 114–119.

[13] Barucci, A., Cucci, C., Franci, M., Loschiavo, M., & Argenti, F. (2021). A deep learning approach to Ancient Egyptian hieroglyphs classification. IEEE Access, PP, 1–1.

[14] Hua, R., & Xu, X. (2019). Intelligent classification on images of Dongba ancient books. The Journal of Engineering, 2019(23), 9039–9042.