

OBI-BENCH: CAN LMMS AID IN STUDY OF ANCIENT SCRIPT ON ORACLE BONES?

Zijian Chen, Tingzhu Chen, Wenjun Zhang, Guangtao Zhai

Shanghai Jiao Tong University

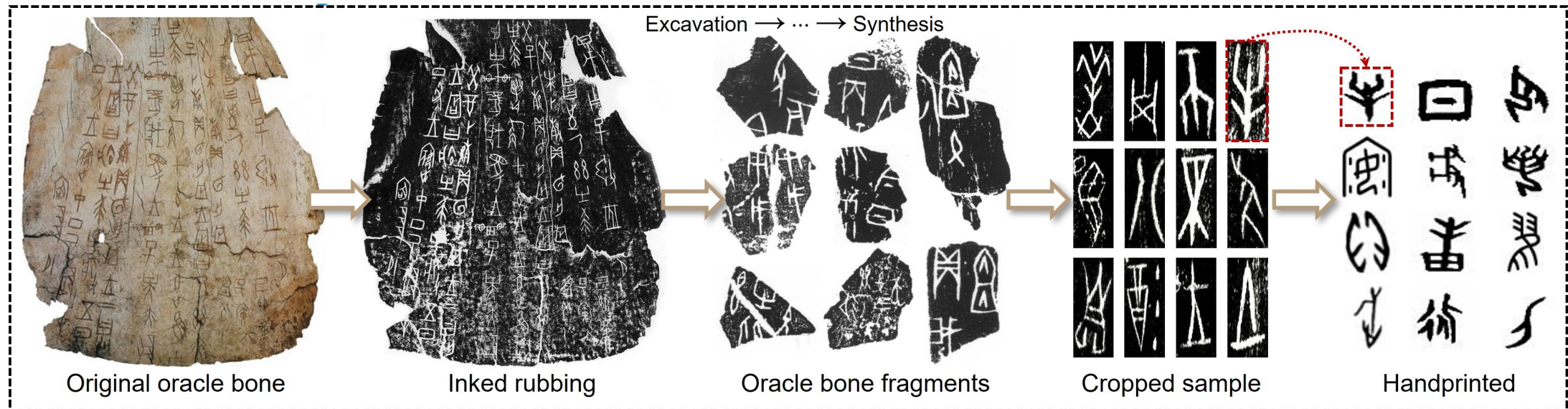
1. Background
2. Construction of OBI-Bench
3. Experiments

What is Oracle Bone Inscription (OBI)?

the earliest known form of Chinese writing, dating back to the late Shang Dynasty (1600–1046 BCE)

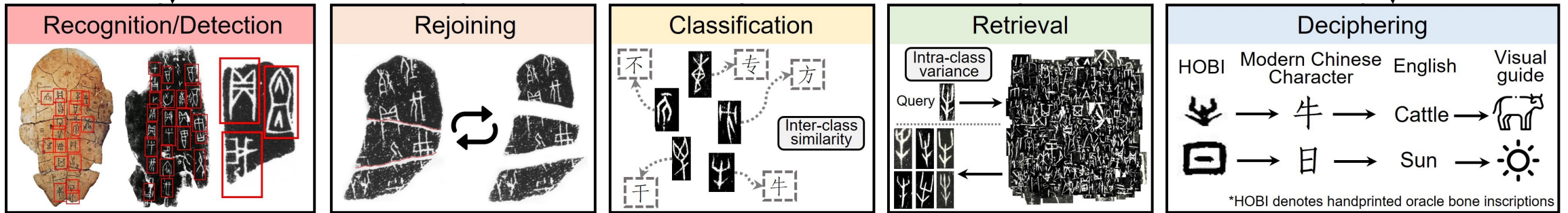
Why we study it?

Ancient history relies on disciplines such as epigraphy—the study of inscribed texts known as inscriptions—for evidence of the thought, language, society and history of past civilizations ^[1]



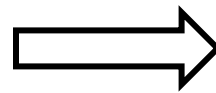
[1] Assael, Yannis, et al. "Restoring and attributing ancient texts using deep neural networks." *Nature* 603.7900 (2022): 280-283.

Five Tasks throughout the OBI Processing



➤ Difficulties in all processing stages

- Lack of standardized coding system
- Lack of literature and corpus
- Lack of data



Making it hard to design a unified and effective AI models for solving these problems

➤ Motivation

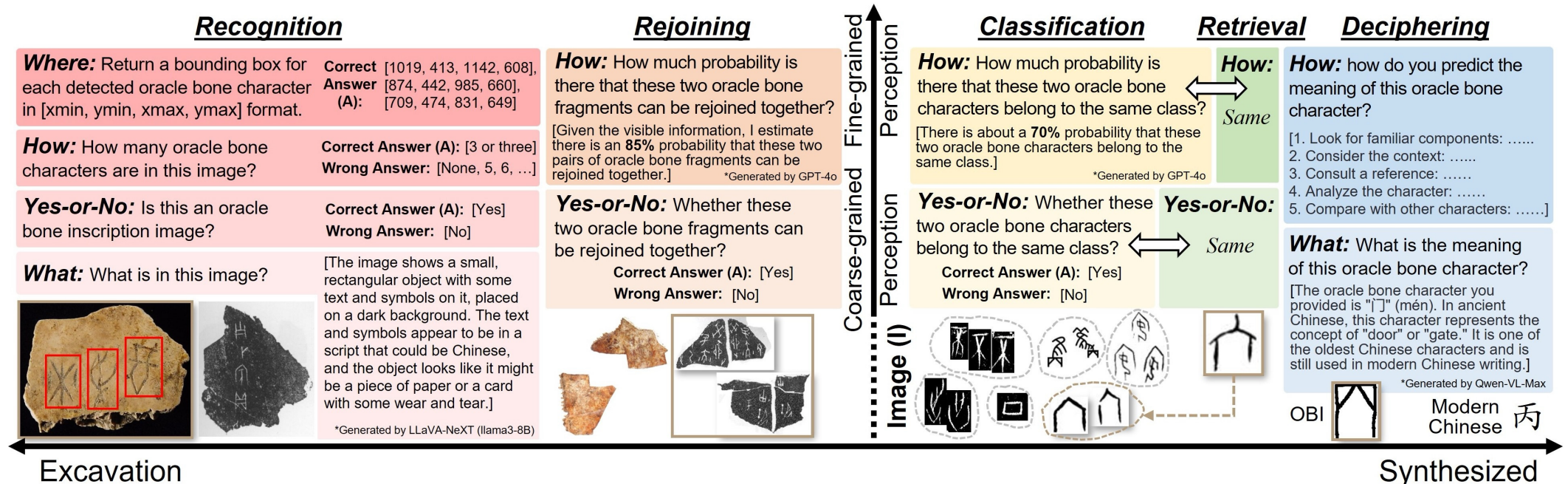
- The **emergent large multi-modal models (LMMs)** have brought opportunities for solving multidisciplinary tasks with powerful **visual perception, understanding, and reasoning** abilities
- The ability of LMMs remains **unclear on fine-grained perception and cognition**, which play significant roles in interpreting ancient texts and its associated tasks on image pre-processing
- LMMs are inherently suited to processing multimodal textual information due to their **natural language-driven characteristics**, which may reduces the knowledge requirements for OBI processing tasks

Can LMMs aid in study of ancient script on oracle bones?

Construction of OBI-Bench

➤ Two Principles

- From Coarse-grained Perception to Fine-grained Perception & Task-oriented Abilities of LMMs.
 - We focus on five major issues in the field of oracle bone inscription research: 1) Recognition; 2) Rejoining; 3) Classification; 4) Retrieval; 5) Deciphering
- Covering Multi-stage Font Appearances.
 - We include all the manifestations of the entire process of OBI processing from excavation (i.e., original oracle bone) to artificial synthesis (i.e., handprinted or computer-generated).



Construction of OBI-Bench

➤ Four Question Types

- **What Question.**
 - In OBI-Bench, they serve as global coarse-grained perception in recognition task (e.g., *What is in this image?*), or unbound the answers in deciphering task (e.g., *What is the meaning of this oracle bone character?*).
- **Yes-or-No Question.**
 - As a fundamental type of judgment, Yes-or-No represents a binary output.
- **How Question.**
 - We also include the How questions to further refine the responses as an extension to Yes-or-No questions. For example, *How many oracle bone characters are in this image?* for the recognition task or *How much probability is there that these two oracle bone characters belong to the same class?*
- **Where Question.**
 - We employ the Where question for the recognition task. For instance, *'Return a bounding box for each detected oracle bone character in [xmin, ymin, xmax, ymax] format to achieve pixel-level OBI anchoring'*

Construction of OBI-Bench

➤ Benchmark Candidates

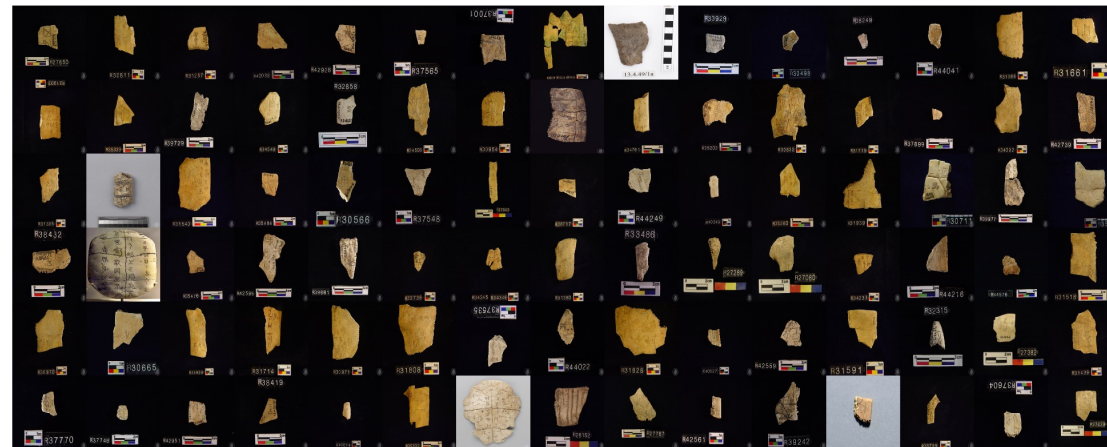
- **6 proprietary LMMs** including Gemini 1.5 Pro, Gemini 1.5 Flash, GPT-4v, GPT-4o, Qwen-VL-Max, and GLM-4v and **17 open-source LMMs**

<i>Date of Release</i> Model Names	Vision Architectures (V)		V→L	Language Architectures (L)	
	Backbone	#Size	Alignment	Backbone	Type
^{24.08} xGcn-MM-instruct-interleave (Xue et al., 2024)	SigLIP-400M	384	Perceiver resampler	Phi-3 mini	decoder-only
^{24.08} mPLUG-Owl3-7B (Ye et al., 2024)	SigLIP-400M	384	MLP projector	Qwen2-7B	decoder-only
^{24.08} MiniCPM-V 2.6-8B (Yao et al., 2024)	SigLIP-400M	224	Perceiver resampler	Qwen2-7B	decoder-only
^{24.08} Moondream2-1.6B (Moondream.ai, 2024)	SigLIP-400M	378	MLP projector	Phi-1.5	decoder-only
^{24.07} InternVL2-Llama3-76B (Chen et al., 2024a)	InternViT-6B	448	MLP projector	Hermes-2-Theta-Llama-3-70B	decoder-only
^{24.07} InternVL2-40B (Chen et al., 2024a)	InternViT-6B	448	MLP projector	Nous-Hermes-2-Yi-34B	decoder-only
^{24.07} InternVL2-8B (Chen et al., 2024a)	InternViT-300M	448	MLP projector	InternLM2.5-7B	decoder-only
^{24.06} GLM-4V-9B (GLM et al., 2024)	EVA2-CLIP-E/14+	224	MLP projector	GLM-4-9B	decoder-only
^{24.05} CogVLM2-LLaMA3-Chat-19B (Wang et al., 2023)	EVA2-CLIP-E	224	MLP projector	Llama-3-8B-Instruct	decoder-only
^{24.05} LLaVA-NcXT-72B (Li et al., 2024a)	CLIP-ViT-L/14	336	MLP projector	Qwen1.5-72B	decoder-only
^{24.05} LLaVA-NcXT-8B (Li et al., 2024a)	CLIP-ViT-L/14	336	MLP projector	Llama-3-8B-Instruct	decoder-only
^{24.05} IDEFICS-2-8B Laurençon et al. (2024)	SigLip-400M	384	MLP projector	Mistral-7B	decoder-only
^{24.03} DeepSeek-VL-7B (Lu et al., 2024)	SAM-B&SigLIP-L	1024&384	MLP projector	DeepSeek-LLM-7B	decoder-only
^{24.01} InternLM-XComposer2-VL-7B (Dong et al., 2024)	CLIP-ViT-L/14	336	Partial LoRA	InternLM2-7B	decoder-only
^{23.10} LLaVA-v1.5-13B (Liu et al., 2024a)	CLIP-ViT-L/14	336	MLP projector	Vicuna-v1.5-13B	decoder-only
^{23.10} LLaVA-v1.5-7B (Liu et al., 2024a)	CLIP-ViT-L/14	336	MLP projector	Vicuna-v1.5-7B	decoder-only
^{23.08} Qwen-VL (Bai et al., 2023)	CLIP-ViT-G/14	448	Cross-attention	Qwen-7B	decoder-only

Experiments

➤ Evaluation on Recognition

- GPT-4o reaches the best performance in terms of the relevance of answers on coarse-grained perception, followed by Qwen-VL-Max
- LMMs are highly effective in handling queries that involve explicitly directed content
- Current LMMs are still not usable for character-level OBI locating and are far from the public-level human

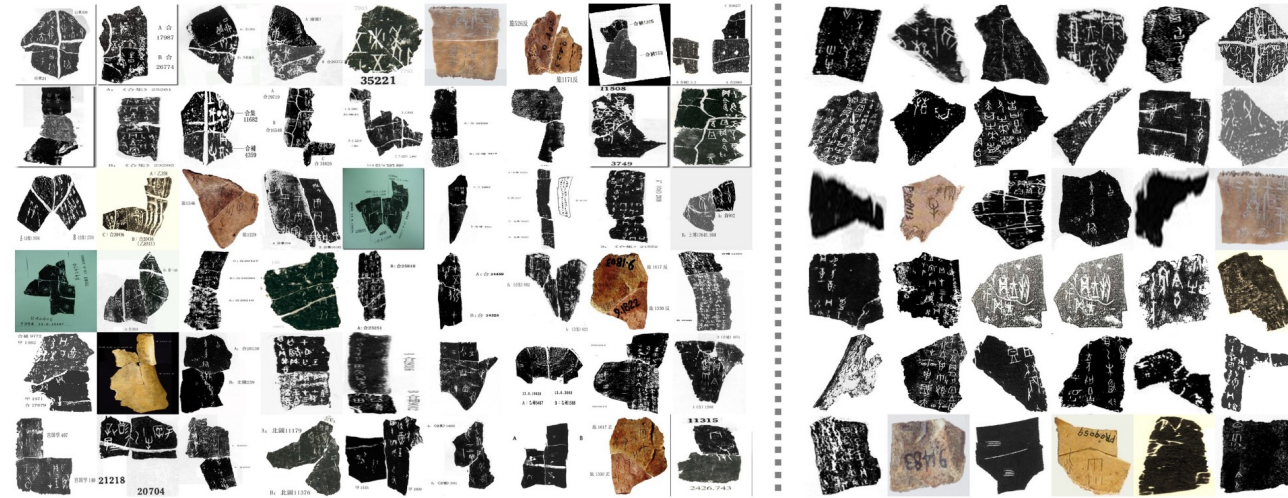


Datasets	O2BR				YinQiWenYuan ^{detection}			
	What↑	Yes-or-No↑	How↓	Where↑	What↑	Yes-or-No↑	How↓	Where↑
LMM (variant)								
HUMAN (public)	0.9364	100%	0.0033	0.9272	0.9189	100%	0.0060	0.8776
Proprietary LMMs:								
GEMINI 1.5 PRO	0.5726	98.75%	0.4932	0.1126	0.3557	99.85%	0.3811	0.0586
GEMINI 1.5 FLASH	0.4123	96.50%	1.7857	0.0962	0.3425	97.35%	0.4522	0.0276
GPT-4v	0.5408	99.30%	0.4223	0.0022	0.3701	99.65%	0.4383	0.0165
GPT-4o (ver. 0806)	0.6114	99.95%	<u>0.4016</u>	0.0038	0.3734	<u>99.90%</u>	0.3458	0.0182
QWEN-VL-MAX (ver. 0809)	<u>0.6071</u>	<u>99.63%</u>	0.4799	0.0086	0.3375	99.55%	0.4843	0.0131
GLM-4V	0.5319	39.17%	0.3681	0.0041	0.3635	52.45%	<u>0.3632</u>	0.0124
Open-source LMMs:								
xGen-MM (Instruct-interleave-4B)	0.5236	81.75%	1.2437	0.0233	0.3669	100%	3.4121	0.0515
mPLUG-Owl3 (Qwen2-7B)	0.3342	99.88%	<u>0.4474</u>	<u>0.0811</u>	0.2505	99.95%	0.6593	0.0527
MiniCPM-V 2.6 (Qwen2-7B)	0.5576	88.75%	0.4829	0.0384	0.3781	99.10%	1.1793	0.0111
Moondream2 (ver. 0728)	0.4818	98.25%	0.6795	0.0400	0.3049	91.30%	<u>0.4436</u>	0.0653
InternVL2-Llama3-76B (Llama3-70B)	0.5833	99.65%	0.4328	0.0976	<u>0.3892</u>	99.75%	0.5344	<u>0.0623</u>
InternVL2-40B (Nous-Hermes2-Yi-34B)	<u>0.5664</u>	98.35%	0.4561	0.0766	0.3637	99.05%	0.6733	0.0487
InternVL2-8B (InternLM2.5-7B)	0.5232	95.00%	0.4618	0.0020	0.3429	97.95%	1.1146	0.0152
GLM-4V-9B (GLM-4-9B)	0.5388	29.50%	0.4825	<10e-4	0.2839	15.70%	0.3934	<10e-4
CogVLM2-Llama3-19B (Llama3-8B)	0.5321	61.00%	0.6928	<10e-4	0.3966	91.75%	0.5568	0.0002
LLaVA-NeXT (Qwen1.5-72B)	0.4846	<u>99.75%</u>	1.1011	0.0445	0.3297	99.75%	1.0561	0.0591
LLaVA-NeXT (Llama3-8B)	0.4764	93.13%	0.5512	0.0001	0.3120	93.90%	0.4268	0.0189
IDEFICS-2-8B (Mistral-7B)	0.3175	95.88%	0.4916	<10e-4	0.2658	95.10%	0.5119	<10e-4
DeepSeek-VL (DeepSeek-LLM-7B)	0.5111	92.75%	0.5657	0.0449	0.3386	98.00%	0.6263	0.0520
InternLM-XComposer2-VL (InternLM2-7B)	0.5106	99.88%	0.6641	0.0049	0.2661	<u>99.95%</u>	1.5231	0.0281
LLaVA-v1.5 (Vicuna-v1.5-13B)	0.4416	99.00%	2.7553	0.0751	0.2875	98.05%	1.7662	0.0493
LLaVA-v1.5 (Vicuna-v1.5-7B)	0.4239	91.88%	0.8861	0.0656	0.2729	81.90%	1.9621	0.0465
Qwen-VL (Qwen-7B)	0.4489	74.13%	1.7416	0.0003	0.3137	86.81%	3.5694	0.0069

Experiments

➤ Evaluation on Rejoining

- The answer in probabilistic form better reflects the differences in visual perception compared to absolute form output
- The number of parameters is roughly proportional to the performance, and newer models tend to exhibit superiority under the same scale of language backbone models
- GPT-4o and Gemini 1.5 Pro reach over 76% in Acc@10 metric, illustrating the possibility of using LMMs to assist the OBI rejoining efforts.
- The overall performance of open-source LMMs is still far away from being truly usable



LMM (variant)	Yes-or-No↑	How↑		
		Acc@1	Acc@5	Acc@10
GEMINI 1.5 PRO	28.53%	24.88%	46.37%	76.67%
GEMINI 1.5 FLASH	22.17%	19.33%	36.34%	66.76%
GPT-4v	27.63%	26.86%	43.75%	73.75%
GPT-4o (ver. 0806)	32.21%	29.13%	48.43%	78.47%
QWEN-VL-MAX (ver. 0809)	21.77%	13.14%	31.68%	56.67%
GLM-4V	5.33%	4.58%	12.19%	21.46%
xGen-MM (Instruct-interleave-4B)	12.08%	3.11%	9.32%	11.18%
mPLUG-Owl3 (Qwen2-7B)	14.48%	3.52%	12.63%	17.60%
MiniCPM-V 2.6 (Qwen2-7B)	13.23%	2.69%	11.18%	16.98%
InternVL2-Llama3-76B (Llama3-70B)	20.13%	5.18%	16.98%	31.68%
InternVL2-40B (Nous-Hermes2-Yi-34B)	16.00%	3.31%	10.97%	24.84%
InternVL2-8B (InternLM2.5-7B)	14.68%	3.73%	10.14%	17.81%
LLaVA-NeXT (Qwen1.5-72B)	17.45%	4.97%	14.29%	27.74%
LLaVA-NeXT (Llama3-8B)	11.41%	3.93%	9.52%	16.77%
IDEFICS-2-8B (Mistral-7B)	7.38%	3.52%	12.42%	14.91%
DeepSeek-VL (DeepSeek-LLM-7B)	9.40%	4.35%	8.90%	11.59%
Qwen-VL (Qwen-7B)	9.86%	4.02%	8.96%	14.08%

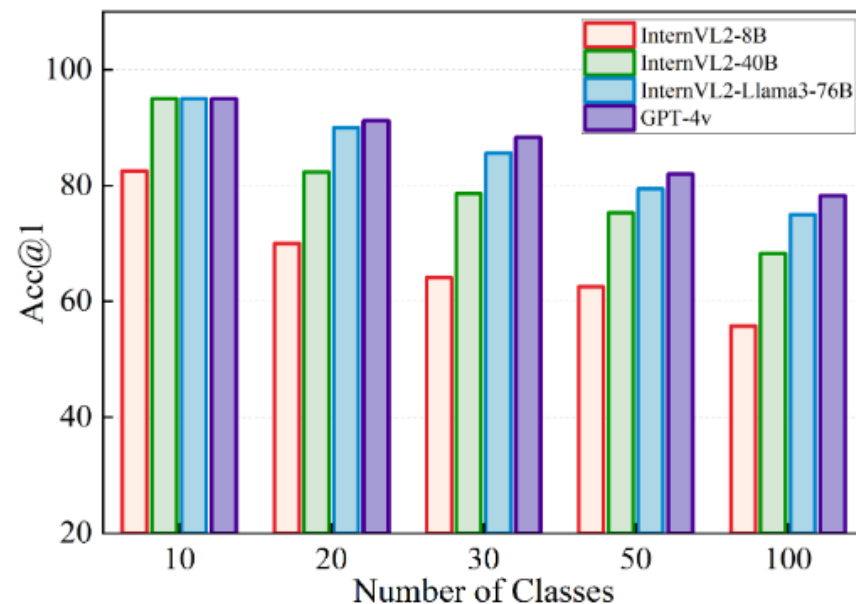
Experiments

➤ Evaluation on Classification

- Image quality of OBI affects the perception of character structure
- Significant performance discrepancies between ‘Yes-or-No’ and ‘How’ queries
- Feasibility of using LMMs for preliminary classification of OBI images.

LMM (variant)	Yes-or-No↑	How↑	
		Acc@1	Acc@5
GEMINI 1.5 PRO	68.75%/68.00%/58.50%	88.75%/85.50%/77.50%	100.0%/100.0%/93.75%
GEMINI 1.5 FLASH	62.50%/60.50%/50.25%	82.00%/82.50%/71.75%	100.0%/99.50%/91.00%
GPT-4v	69.50%/66.00%/57.75%	86.75%/88.25%/70.75%	100.0%/100.0%/91.75%
GPT-4o (ver. 0806)	72.75%/74.50%/62.50%	89.75%/90.25%/75.50%	100.0%/100.0%/93.50%
QWEN-VL-MAX (ver. 0809)	64.25%/65.75%/55.00%	85.00%/88.75%/69.25%	100.0%/98.75%/89.75%
GLM-4V	35.50%/29.75%/34.50%	71.00%/68.50%/54.25%	83.50%/85.75%/77.75%
xGen-MM (Instruct-interleave-4B)	36.25%/37.75%/35.75%	46.75%/48.00%/43.50%	57.75%/55.00%/49.00%
mPLUG-Owl3 (Qwen2-7B)	39.25%/39.75%/38.75%	44.75%/46.25%/43.75%	56.25%/52.50%/53.50%
MiniCPM-V 2.6 (Qwen2-7B)	40.75%/42.50%/38.50%	45.75%/45.00%/42.75%	56.75%/55.75%/51.75%
InternVL2-Llama3-76B (Llama3-70B)	44.75%/47.50%/43.25%	53.75%/55.00%/50.75%	69.75%/69.00%/66.75%
InternVL2-40B (Nous-Hermes2-Yi-34B)	42.75%/43.50%/40.25%	49.75%/50.00%/48.75%	63.75%/62.25%/59.75%
InternVL2-8B (InternLM2.5-7B)	42.25%/41.75%/38.75%	47.75%/49.00%/47.75%	59.75%/59.00%/56.50%
LLaVA-NeXT (Qwen1.5-72B)	46.25%/45.50%/41.00%	51.75%/52.00%/49.00%	66.75%/67.75%/64.25%
LLaVA-NeXT (Llama3-8B)	44.00%/42.00%/38.75%	46.75%/46.25%/42.75%	53.75%/56.75%/54.75%
IDEFICS-2-8B (Mistral-7B)	44.75%/44.50%/39.75%	46.75%/46.50%/44.00%	58.75%/58.75%/53.75%
DeepSeek-VL (DeepSeek-LLM-7B)	42.25%/40.75%/35.75%	47.00%/47.25%/45.25%	56.75%/58.25%/58.50%
Qwen-VL (Qwen-7B)	44.25%/42.00%/38.75%	48.00%/51.00%/44.75%	61.25%/63.50%/61.25%

- Overall performance improvement → the source-mixed test set increases the inter-class differences.
- The accuracy of these models decreases as the number of classes increases.
- LMMs with larger parameter sizes tend to have better robustness



Experiments

➤ Evaluation on Retrieval


LMM (variant)	Yes-or-No↑	How↑			
	$\overline{mAP@ Yes }$	$\overline{Recall@1}$	$\overline{Recall@3}$	$\overline{Recall@10}$	$\overline{mAP@5}$
GEMINI 1.5 PRO	<u>0.4316/0.5734</u>	<u>0.225/0.250</u>	0.671/0.688	1.000/1.000	0.644/0.76
GEMINI 1.5 FLASH	0.3876/0.5344	0.195/ <u>0.225</u>	0.635/0.644	<u>0.975/1.000</u>	0.562/0.74
GPT-4v	0.4228/0.5680	0.205/ <u>0.225</u>	0.650/0.676	1.000/1.000	0.624/0.74
GPT-4o (ver. 0806)	0.4550/0.6122	0.235/0.250	0.686/0.706	1.000/1.000	0.688/0.80
QWEN-VL-MAX (ver. 0809)	0.4223/0.5716	0.190/ <u>0.225</u>	0.621/0.638	1.000/1.000	<u>0.664/0.78</u>
GLM-4V	0.3018/0.3867	0.125/0.175	0.495/0.613	0.925/ 1.000	0.520/0.72
xGen-MM (Instruct-interleave-4B)	<u>0.2668/0.3423</u>	<u>0.085/0.200</u>	<u>0.336/0.569</u>	<u>0.855/0.975</u>	<u>0.366/0.66</u>
mPLUG-Owl3 (Qwen2-7B)	0.2814/0.3550	0.075/ <u>0.225</u>	0.371/0.588	0.870/0.950	0.384/0.62
MiniCPM-V 2.6 (Qwen2-7B)	0.2863/0.3575	0.070/ <u>0.225</u>	0.361/0.581	0.885/0.925	0.368/0.66
InternVL2-Llama3-76B (Llama3-70B)	0.3557/0.4268	<u>0.150/0.250</u>	0.460/0.675	1.000/1.000	<u>0.522/0.72</u>
InternVL2-40B (Nous-Hermes2-Yi-34B)	0.3260/0.3914	0.125/ 0.250	0.445/0.656	1.000/1.000	0.480/0.68
InternVL2-8B (InternLM2.5-7B)	0.2844/0.3623	0.095/ <u>0.225</u>	0.374/0.650	0.925/ 1.000	0.420/0.68
LLaVA-NeXT (Qwen1.5-72B)	<u>0.3478/0.4143</u>	0.155/0.250	0.468/0.669	1.000/1.000	0.562/0.70
LLaVA-NeXT (Llama3-8B)	0.2793/0.3605	0.075/ <u>0.225</u>	0.358/0.606	0.925/ 1.000	0.348/0.60
IDEFICS-2-8B (Mistral-7B)	0.2844/0.3602	0.090/0.200	0.354/0.619	0.875/ 1.000	0.402/0.66
DeepSeek-VL (DeepSeek-LLM-7B)	0.2867/0.3450	0.105/ <u>0.225</u>	0.343/0.600	<u>0.930/1.000</u>	0.416/0.64
Qwen-VL (Qwen-7B)	0.2883/0.3528	0.080/ <u>0.225</u>	0.345/0.588	0.920/0.925	0.422/0.66

Experiments

➤ Evaluation on Deciphering

- The deciphering performance on common characters is better than that on rare characters.
- The deciphering results on pictograph are more accurate than ideogram
- LMMs are impeded in distinguishing and deciphering component-level variants
- GPT-4o, GPT-4v, and Gemini 1.5 Pro have approached or exceeded public-level humans in some scenarios

Datasets	HUST-OBS			EVOBC		OBI Component 20	Average
	LMM (variant)	Tier-1	Tier-2	Tier-3	pictograph	ideogram	
HUMAN (public)		0.4507	0.3884	0.3437	0.4966	0.3627	0.3872
GEMINI 1.5 PRO		0.3766	0.3834	0.3589	0.4226	0.3696	0.3810
GEMINI 1.5 FLASH		0.3545	0.3829	0.3567	0.3661	0.3595	0.3641
GPT-4v		0.3764	0.3808	0.3596	0.4424	0.3551	0.3867
GPT-4o (ver. 0806)		0.3891	0.3510	0.3660	0.4535	0.3893	0.3876
QWEN-VL-MAX (ver. 0809)		0.2345	0.2273	0.2322	0.2565	0.2533	0.2471
GLM-4V		0.2180	0.1638	0.2122	0.2353	0.2451	0.2071
xGen-MM (Instruct-interleave-4B)		0.1174	0.1332	0.1369	0.1526	0.1268	0.1283
mPLUG-Owl3 (Qwen2-7B)		0.1609	0.1822	0.1897	0.2036	0.1751	0.1633
MiniCPM-V 2.6 (Qwen2-7B)		0.1411	0.1694	0.1584	0.1703	0.1529	0.1471
Moondream2 (ver. 0728)		0.1129	0.1171	0.1063	0.1297	0.1176	0.1165
InternVL2-Llama3-76B (Llama3-70B)		0.2324	0.2139	0.2156	0.2355	0.2123	0.2163
InternVL2-40B (Nous-Hermes2-Yi-34B)		0.1966	0.1917	0.1883	0.2007	0.1902	0.1842
InternVL2-8B (InternLM2.5-7B)		0.1676	0.1739	0.1340	0.1746	0.1517	0.1615
GLM-4V-9B (GLM-4-9B)		0.1291	0.0821	0.0478	0.0835	0.0909	0.0789
CogVLM2-Llama3-19B (Llama3-8B)		0.1096	0.0873	0.0852	0.1024	0.1079	0.1019
LLaVA-NeXT (Qwen1.5-72B)		0.1501	0.1647	0.1372	0.1468	0.1123	0.1329
LLaVA-NeXT (Llama3-8B)		0.0893	0.0873	0.0892	0.0806	0.0813	0.0763
IDEFICS-2-8B (Mistral-7B)		0.1639	0.1252	0.0985	0.1601	0.2189	0.1531
DeepSeek-VL (DeepSeek-LLM-7B)		0.0458	0.0628	0.0504	0.0324	0.0593	0.0573
InternLM-XComposer2-VL (InternLM2-7B)		0.2963	0.3101	0.2660	0.2650	0.1822	0.2304
LLaVA-v1.5 (Vicuna-v1.5-13B)		0.1368	0.1669	0.1629	0.1128	0.1214	0.1363
LLaVA-v1.5 (Vicuna-v1.5-7B)		0.0677	0.1165	0.0906	0.0924	0.0887	0.1008
Qwen-VL (Qwen-7B)		0.1145	0.0912	0.0757	0.1080	0.0635	0.0879



OBI

Modern Chinese: 王


Golden Description: This oracle bone character denotes "king", signifying a ruler or monarch in ancient.

GPT-4o BERTScore=0.1946 ❌

The Oracle Bone character in the image represents the number "five" (五) in ancient Chinese script. This script was used during the Shang dynasty and is one of the earliest forms of Chinese writing.

Qwen-VL-Max 0.2829 ❌

The oracle bone character you provided is "大" (dà). In ancient Chinese, this character represents the concept of "big" or "large." It is one of the oldest Chinese characters and is still used in modern Chinese writing.



OBI

Modern Chinese: 子


Golden Description: This oracle bone character denotes "child", typically referring to a young offspring.

GPT-4o BERTScore=0.2345 ❌

The Oracle Bone character in the image represents the word "mouth" in ancient Chinese script. This character is a pictograph depicting a human mouth, which has been preserved in modern Chinese writing as 口.

Qwen-VL-Max 0.3143 ❌

The oracle bone character you provided is 人 (rén). In ancient, this character represents the concept of "person" or "people." It is one of the oldest Chinese characters and is still used in modern Chinese writing.



OBI

Modern Chinese: 雨


Golden Description: This oracle bone character denotes "rain". The upper part symbolizes the sky or clouds, while the lower part represents raindrops.

GPT-4o BERTScore=0.5885 ✅

This Oracle Bone character represents rain (雨). The top portion of the character resembles the sky or clouds, while the vertical lines below symbolize raindrops falling to the ground.

Qwen-VL-Max 0.2295 ❌

The oracle bone character you provided is 三 (sān). In ancient Chinese, this character represents the number "three." It is one of the oldest Chinese characters and is still used in modern Chinese writing.



OBI

Modern Chinese: 牛

Golden Description: This Oracle Bone character denotes "cow" or "ox," representing an important domesticated animal in ancient Chinese society.

GPT-4o BERTScore=0.2677 ❌

The Oracle Bone character in the image represents the word "branch" or "forked" (支). It visually depicts something branching out or splitting, which is consistent with the modern Chinese character 支.

Qwen-VL-Max 0.3335 ❌

The oracle bone character you provided is 木 (mù). In ancient Chinese, this character represents the concept of "tree" or "wood." It is one of the oldest Chinese characters and is still used in modern Chinese writing.

Thanks !

Please contact Zijian Chen for any questions: zijian.chen@sjtu.edu.cn