On Path to Multimodal Historical Reasoning: HistBench and HistAgent

Jiahao Qiu[*1], Fulian Xiao[*2], Yimin Wang[*3,4], Yuchen Mao[*4], Yijia Chen[*5], Xinzhe Juan[3,4], Siran Wang[2], Xuan Qi[6], Tongcheng Zhang[4], Zixin Yao[7], Jiacheng Guo[1], Yifu Lu[1], Charles Argon[8], Jundi Cui[2], Daixin Chen[5], Junran Zhou[2], Shuyao Zhou[1], Zhanpeng Zhou[4], Ling Yang[1], Shilong Liu[1], Hongru Wang[9], Kaixuan Huang[1], Xun Jiang[10,11], Xi Gao[†2], Mengdi Wang[†1]

[1] Princeton [2] Fudan (History) [3] Michigan [4] SJTU [5] Fudan (Philosophy) [6] Tsinghua [7] Columbia [8] Princeton (History) [9] Edinburgh [10] TCCI [11] Theta Health

[*] Equal Contribution [†] Corresponding Authors

ICML 2026

Dataset Contributors

We gratefully acknowledge the contributors who authored and reviewed HistBench questions:

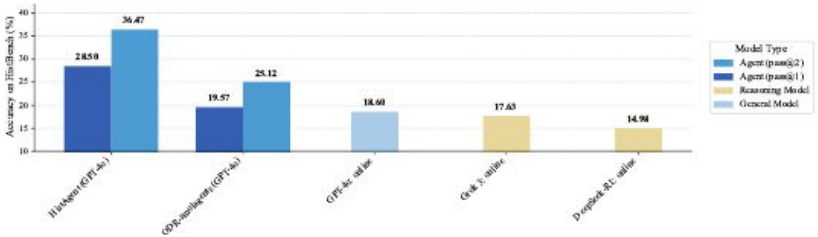
Yuming Cao, Yue Chen, Yunfei Chen, Zhengyi Chen, Ruowei Dai, Mengqiu Deng, Jiye Fu, Yunting Gu, Zijie Guan, Zirui Huang, Xiaoyan Ji, Yumeng Jiang, Delong Kong, Haolong Li, Jiaqi Li, Ruipeng Li, Tianze Li, Zhuoran Li, Haixia Lian, Mengyue Lin, Xudong Liu, Jiayi Lu, Jinghan Lu, Wanyu Luo, Ziyue Luo, Zihao Pu, Zhi Qiao, Ruihuan Ren, Liang Wan, Ruixiang Wang, Tianhui Wang, Yang Wang, Zeyu Wang, Zihua Wang, Yujia Wu, Zhaoyi Wu, Hao Xin, Weiao Xing, Ruojun Xiong, Weijie Xu, Yao Shu, Yao Xiao, Xiaorui Yang, Yuchen Yang, Nan Yi, Jiadong Yu, Yangyuxuan Yu, Huiting Zeng, Danni Zhang, Yunjie Zhang, Zhaoyu Zhang, Zhiheng Zhang, Xiaofeng Zheng, Peirong Zhou, Linyan Zhong, Xiaoyin Zong, Ying Zhao, Zhenxin Chen, Lin Ding, Xiaoyu Gao, Bingbing Gong, Yichao Li, Yang Liao, Guang Ma, Tianyuan Ma, Xinrui Sun, Tianyi Wang, Han Xia, Ruobing Xian, Gen Ye, Tengfei Yu, Wentao Zhang, Yuxi Wang.

Motivation & The Gap

LLM agents excel in science, code, math... but the humanities remain underexplored.

Why **history** as the first humanities testbed?

- Incomplete, heterogeneous evidence
- Cross-lingual, cross-modal, cross-cultural
- Large, well-annotated corpora

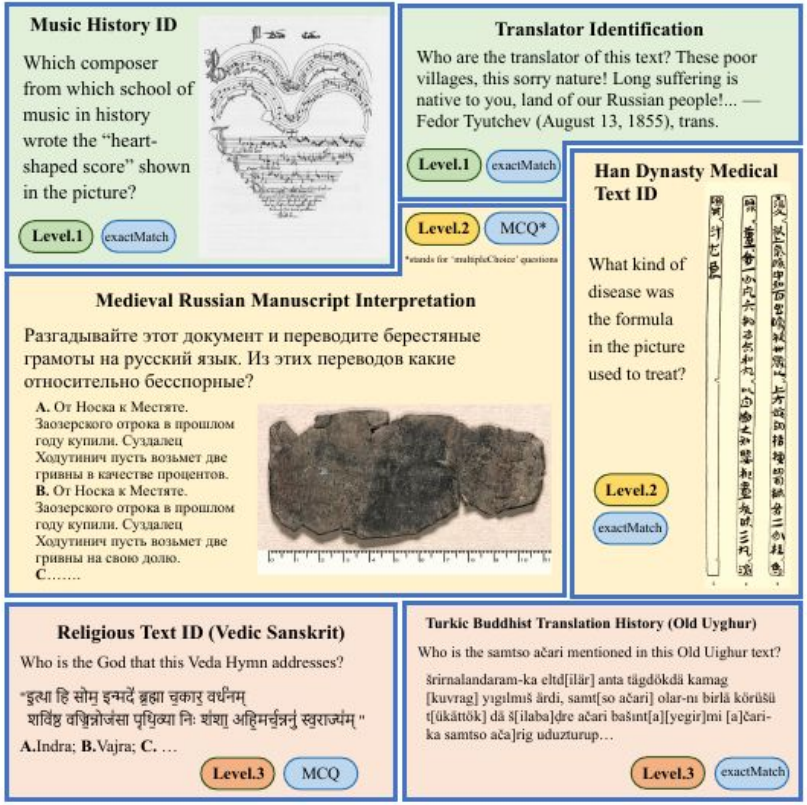


Performance of LLMs and Agents on HistBench

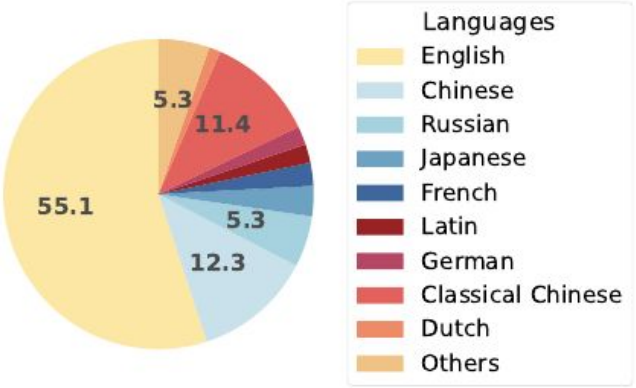
No existing benchmark systematically tests multilingual sources, multimodal evidence, and evidence-grounded historical reasoning.

- GAIA, MMLU-Pro — no disciplinary depth
- HLE — only **56** history questions
- HiST-LLM — tests *knowledge*, not *research*

HistBench: Dataset Composition & Diversity



(a) Six representative questions (Levels 1–3)



(b) 29 modern & ancient languages

(c) Geographic coverage — all inhabited continents

HistBench: Difficulty Stratified by Six Expert Dimensions

	Rarity of Source Knowledge	Linguistic Complexity	Format Heterogeneity	Perceptual Accessibility	Interdisciplinarity	Reasoning Complexity
Level 1: Basic 	Common scholarly literature and textbooks	English or modern Chinese	Single-source, well-structured documents	Clear print or modern editions	Intra-disciplinary (within history)	Factual recall, single-step inference
Level 2: Intermediate 	Specialized databases or unpublished records	French, German, Spanish	Text-image combinations	Cursive handwriting, early print, marginalia	Bi-disciplinary (e.g., history + economics)	Multi-sentence interpretation, basic causality
Level 3: Challenging 	Rare, obscure, or marginal archival materials	Dead languages or rarely taught scripts (e.g., Latin, Syriac)	Multi-modal, cross-format corpora	Fragmentary inscriptions, damaged scans, nonstandard script	Multi-disciplinary (e.g., history + archaeology + theology)	Deep reasoning across sources with conditional constraints

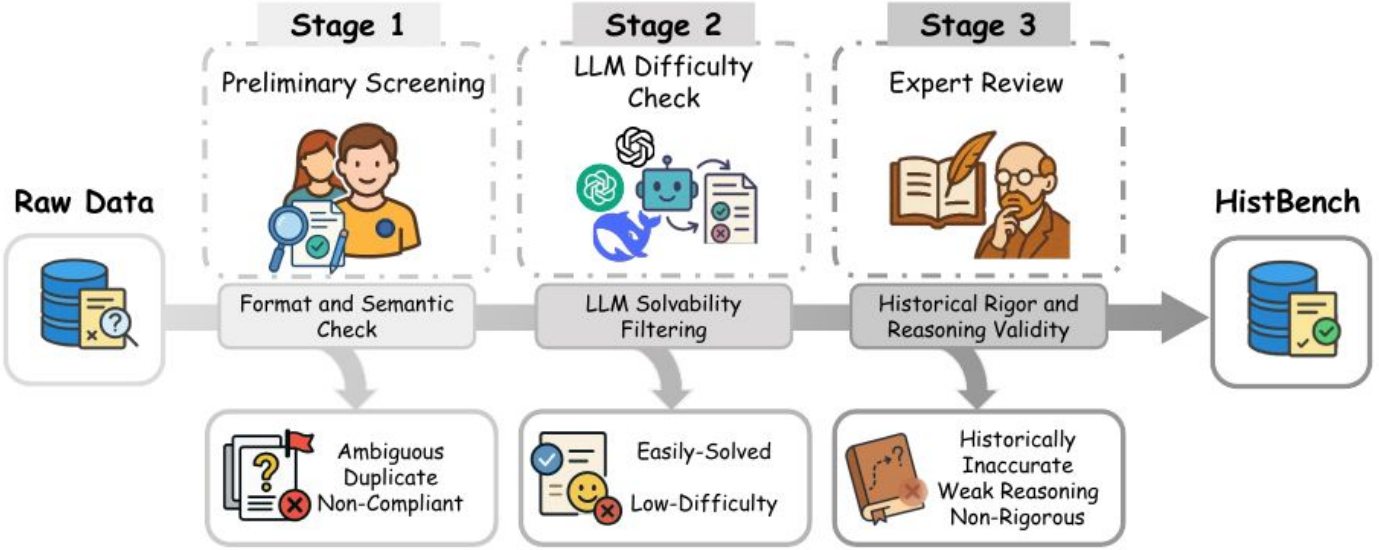
Difficulty

Easy

Hard

Three difficulty levels — not by model performance, but by *human historians*: **L1 (Basic)** drafted by trained RAs · **L2 (Intermediate)** by graduate researchers · **L3 (Challenging)** by senior scholars.

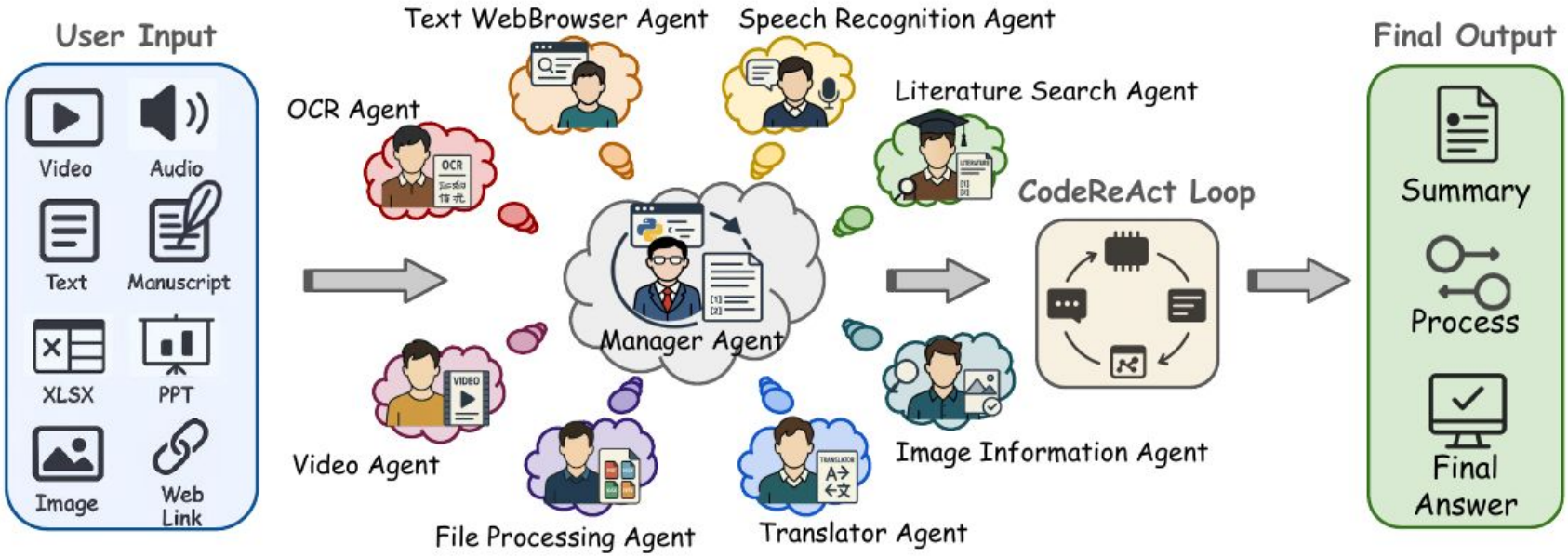
HistBench: Three-Stage Curation Pipeline



1. **Preliminary screening** — clarity, completeness, redundancy
2. **LLM difficulty check** — discard if ≥ 2 of {GPT-4o, GPT-4-mini, DeepSeek-R1} solve without sources
3. **Expert validation** — historians review evidentiary grounding & methodological rigor

$$\Rightarrow \textbf{1034 candidates} \rightarrow \textbf{720 retained} \rightarrow \textbf{414 final.}$$

HistAgent: Manager–Specialist Architecture



A central **Manager Agent** runs a CodeAct loop and dispatches **specialist agents**: handwriting **OCR** (Transkribus), multilingual **translation**, reverse **image search**, **scholarly retrieval** (Google Scholar, Springer), plus file parsing, speech, and video tools.

Main Result: HistBench (414 Q)

Agent / Model		Level 1	Level 2	Level 3	Average
HistAgent (GPT-4o)	pass@1	28.92	26.16	32.89	28.50
	pass@2	**36.14**	**35.47**	**39.47**	**36.47**
ODR-smolagents (GPT-4o)	pass@1	16.27	21.51	22.37	19.57
	pass@2	20.48	28.49	27.63	25.12
DeepSeek-R1:online		11.45	19.77	11.84	14.98
GPT-4o:online		13.86	21.51	22.37	18.60
o4-mini-high:online		28.92	30.81	31.58	30.19
Grok 3:online		13.25	19.77	22.37	17.63

+11.35 pts pass@2 over ODR-smolagents (same GPT-4o backbone).
Competitive with much stronger closed models (o4-mini-high) despite a weaker backbone.

Generalization: HLE-History & GAIA

HLE History Subset (56 Q)

System	p@1	p@2	p@3
HistAgent (GPT-4o)	**28.57**	**39.29**	**42.86**
ODR-smolagents (GPT-4o)	17.86	25.00	28.57
GPT-4o + web search	8.93	19.64	25.00

3× the pass@1 of plain GPT-4o

GAIA Validation (165 Q)

Agent	Avg	L1	L2	L3
HistAgent (Claude-3.7)	**60.00**	69.81	61.63	34.62
ODR-smolagents (o1)	55.15	67.92	53.49	34.62

Domain specialization **does not hurt** general competence

Gains generalize beyond HistBench — modality breakdown shows the **multimodal** advantage comes from **OCR, translation, and reverse image search**.

Take-aways

- **HistBench** — the first benchmark systematically targeting historical reasoning: 414 questions, 29 languages, 36 subfields, expert-stratified difficulty.
- **HistAgent** — a manager–specialist agent with history-oriented tools (handwriting OCR, scholarly retrieval, translation, image analysis).
- **Result** — **36.47% pass@2** on HistBench and **60.00% pass@1** on GAIA with a GPT-4o backbone.

Targeted **tool & reasoning design** can partly offset model scale — pointing toward affordable, domain-aware agents for the humanities.

Thank you!

Code: `github.com/CharlesQ9/HistAgent`