arXiv:2505.20246v3 [cs.AI] 19 Jun 2025

ON PATH TO MULTIMODAL HISTORICAL REASONING: HISTBENCH AND HISTAGENT

Organizing Team

Jiahao Qiu*[1], Fulian Xiao*[2], Yimin Wang*[3], Yuchen Mao*[4], Yijia Chen*[5], Xinzhe Juan[3], Shu Zhang[11], 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]**AI Lab, Princeton University** [2]**Department of History, Fudan University** [3]**University of Michigan** [4]**Shanghai Jiao Tong University** [5]**School of Philosophy, Fudan University** [6]**IIIS, Tsinghua University** [7]**Department of Philosophy, Columbia University** [8]**Department of History, Princeton University** [9]**The Chinese University of Hong Kong** [10]**Tianqiao and Chrissy Chen Institute** [11]**Theta Health Inc.**

Dataset Contributors

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

ABSTRACT

Recent advances in large language models (LLMs) have led to remarkable progress across various domains, yet their capabilities in the humanities, particularly history, remain underexplored. Historical reasoning poses unique challenges for AI, involving multimodal source interpretation, temporal inference, and cross-linguistic analysis. While general-purpose agents perform well on many existing benchmarks, they lack the domain-specific expertise required to engage with historical materials and historical questions. To address this gap, we introduce HistBench, a new benchmark of 414 high-quality questions designed to evaluate AI's capacity for historical reasoning and authored by more than 40 expert contributors. The tasks span a wide range of historical problems—from factual retrieval based on primary sources to interpretive analysis of manuscripts and images, to interdisciplinary challenges involving archaeology, linguistics, or cultural history. Furthermore, the benchmark dataset spans 29 ancient and modern languages and covers a wide range of historical periods and world regions. Finding the poor performance of LLMs and other agents on HistBench, we further present HistAgent, a history-specific agent equipped with carefully designed tools for OCR, translation, archival search, and image understanding in History. On HistBench, HistAgent based on GPT-4o achieves an accuracy of 27.54% pass@1 and 36.47% pass@2, significantly outperforming LLMs with online search and generalist agents, including GPT-4o (18.60%), DeepSeek-R1(14.49%), Grok 3(17.63%) and Open Deep Research-smolagents(20.29% pass@1 and 25.12% pass@2). These results highlight the limitations of existing LLMs and generalist agents and demonstrate the advantages of HistAgent for historical reasoning. Notably, HistAgent also achieves 60.00% pass@1 accuracy on GAIA, showing that domain-specific customization doesn't hinder HistAgent's competitive performance on real-world general tasks.

Code: `https://github.com/CharlesQ9/HistAgent`
Dataset: `https://huggingface.co/datasets/jiahaoq/HistBench`

* Equal contribution.
† Correspondence to: `gaoxi@fudan.edu.cn`, `mengdiw@princeton.edu`.

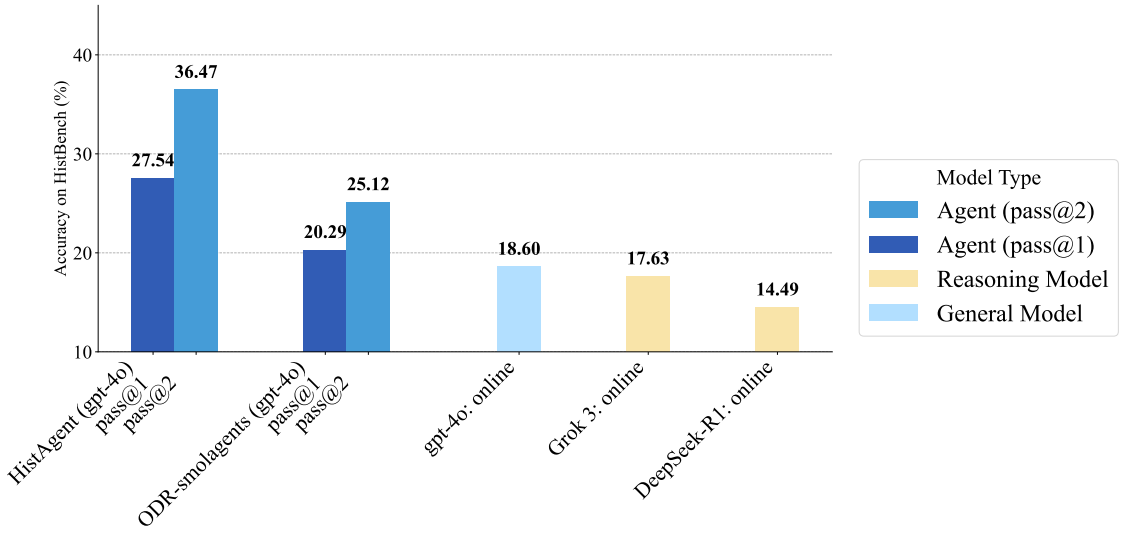


Figure 1: Performance of LLMs and Agents on HistBench.

1 Introduction

Recent advances in large language models (LLMs) have led to AI agents capable of impressive performance across a range of complex tasks. Although significant progress has been made in general-purpose and AI4Science agents, such as OpenAI's Deep Research [1] and agents for chemistry and biology [2, 3], the humanities remain critically underexplored. This imbalance risks narrowing both the scope of artificial intelligence and our understanding of intelligence itself. If AI is to become truly general, it must also engage with the complex and underexplored challenges posed by the humanities.

Among the humanities, history holds a uniquely central role. It addresses fundamental questions of human identity, continuity, and transformation, while exemplifying the interpretive complexity of humanistic scholarship. Historical reasoning requires navigating incomplete, heterogeneous sources—ranging from manuscripts and inscriptions to maps and visual records—and demands cross-linguistic, cross-modal, and cross-cultural interpretation. Unlike factual recall, it involves temporal reasoning, contextualization, and the reconciliation of conflicting narratives. As both epistemically rich and methodologically demanding, history encapsulates many of the core challenges that characterize reasoning in the humanities. In addition to its disciplinary significance, history is also the most computationally tractable field among the humanities. It generates and consumes large-scale textual and visual data, far exceeding most other humanities domains in volume, structure, and annotation potential. Moreover, historical research is inherently global and multilingual, shaped by regionally specific contexts. These features make history an ideal testbed for evaluating AI systems under real-world challenges such as cross-lingual retrieval, multi-modal reasoning, and cultural alignment. For these reasons, history is not only methodologically appropriate but also strategically positioned as the first point of engagement for AI-humanities research.

Despite the proliferation of AI benchmarks, there remains no rigorous evaluation suite specifically designed to assess historical reasoning. Most existing benchmarks either focus on general capabilities or are tailored to domains like science and mathematics. For example, the GAIA benchmark evaluates real-world tasks across broad domains, but does not target any single discipline in depth [4]. Humanity's Last Exam (HLE), a large-scale human-written test covering over 2,500 questions, includes only 56 history-related problems [5]—far too few to evaluate the interpretive, evidentiary, and temporal reasoning skills central to historical reasoning. Meanwhile, discipline-specific benchmarks such as PhyBench [6] demonstrate the importance of domain-adapted evaluation in fields like physics, but no equivalent exists for history. As a result, there is currently no benchmark that captures the unique methodological demands of historical research and historical reasoning, leaving a critical gap in evaluating LLM and agent performance in this foundational area of the humanities.

To address this gap, we introduce HistBench, the first comprehensive benchmark dedicated to evaluating historical reasoning in AI systems. The dataset contains 414 high-quality and carefully-reviewed questions, contributed by history professors, PhD students, and research assistants. Each question simulates a realistic research task and is annotated with difficulty level, question type, and reasoning dimension. The tasks span a wide range of historical problems—from

factual retrieval based on primary sources to interpretive analysis of manuscripts and images, to interdisciplinary challenges involving archaeology, linguistics, or cultural history. HistBench questions are categorized into three difficulty levels—Level 1 (Basic), Level 2 (Intermediate), and Level 3 (Challenging)—based on six structured criteria: rarity of source knowledge, linguistic complexity, format heterogeneity, perceptual accessibility, interdisciplinary scope, and reasoning depth. The benchmark covers 29 languages and all major world regions, and all questions are carefully designed to resist resolution through retrieval-based answering or shallow prompting. By combining topical breadth with rigorous, source-grounded construction, HistBench enables domain-specific evaluation of LLMs under expert-defined historical reasoning constraints.

With the introduction of HistBench, we can now systematically evaluate how well LLMs and generalist agents perform on historical tasks. Our findings reveal a consistent gap: current large language models and agents lack the capabilities required for historical reasoning. Unlike other tasks, historical research involves interpreting fragmented, multimodal sources across languages, formats, and time periods. However, today's models and agents fall short in meeting the demand for historical reasoning. Their OCR tools fail on handwritten manuscripts, damaged archives, and early print formats; their translation capabilities are limited to high-resource modern languages, with poor support for classical and minority languages; and their visual understanding of historical materials—such as annotated maps, marginalia, or iconography—is minimal. These limitations are not only technical but methodological: historical analysis demands multi-step reasoning, cross-source synthesis, and epistemic grounding that current general-purpose systems do not support. This performance gap underscores the need for dedicated agents equipped with domain-specific tools and workflows aligned with historical practice.

To overcome the limitations of large language models and generalist agents, we introduce HistAgent, a domain-specialized AI agent tailored for historical reasoning. Built on top of an advanced LLM, HistAgent integrates a set of modular tools designed to address the epistemic and technical challenges unique to historical inquiry. These include OCR modules for handwritten manuscripts (Transkribus, Asian-script OCR), multilingual translators with provenance preservation, reverse image search for historical visuals, and peer-reviewed literature retrieval mimicking history researchers. HistAgent spans multiple modalities—text, image, audio, manuscript, video—and applies source-aware workflows for extraction, parsing, and reasoning.

On HistBench, HistAgent based on GPT-4o achieves an accuracy of 27.54% pass@1 and 36.47% pass@2, significantly outperforming LLMs with online search and generalist agents, including GPT-4o (18.60%), DeepSeek-R1 (14.49%), Grok 3(17.63%) and open Deep Research by smolagents(20.29% pass@1 and 25.12% pass@2). These results highlight the limitations of existing LLMs and generalist agents and demonstrate the advantages of HistAgent for historical reasoning. Notably, HistAgent also achieves 60.00% pass@1 accuracy on the GAIA benchmark, showing that domain-specific customization doesn't hinder HistAgent's competitive performance on real-world general tasks.

In summary, this paper makes the following contributions:

- **HistBench**: We present HistBench, the first large-scale and comprehensive benchmark involving 414 high-quality and carefully-reviewed questions for evaluating historical reasoning in AI systems.
- **HistAgent**: We present HistAgent, an agent specialized for the history domain. HistAgent integrates multi-modal perception, web browsing, and a suite of modular tools—including OCR for manuscripts, image provenance retrieval, multilingual translation, literature parsing, and document analysis—into a cohesive framework for historical research. It is designed to mirror the interpretive workflow of human historians, enabling document-level reasoning, multimodal synthesis, and context-aware analysis across languages and source formats. By combining general LLM capabilities with domain-specific customization, HistAgent overcomes key limitations of existing LLMs and generalist agents in handling historical tasks while maintaining competitive performance on general tasks.
- **Comprehensive Empirical Validation**: We provide an extensive evaluation of HistAgent on both domain-specific and general benchmarks. HistAgent significantly outperforms base LLMs with online search on HistBench and the history subset of Humanity's Last Exam, demonstrating superior historical reasoning. Furthermore, HistAgent achieves high performance on real-world general-purpose benchmarks like GAIA, highlighting that a history-focused agent can maintain broad capabilities and excel in complex, multi-step reasoning tasks beyond its core domain.

By introducing HistBench and HistAgent, this work lays a concrete foundation for the path toward multi-modal historical reasoning in AI. Our benchmark captures the complexity of historical inquiry across languages, media, and reasoning styles, while our agent demonstrates how agents can begin to emulate the interpretive workflows of human historians. We hope this effort not only advances AI for history, but also inspires broader exploration of LLMs and domain-specific agents across other underrepresented areas of human knowledge and marks a broader step toward AI for the Humanities.

2 Related Works

2.1 Generalist Agent

Generalist agent is for solving general real-world tasks such as GAIA[4]. OMNE [7] introduces a column-structured long-term memory that agents update during inference to refine policies without retraining. AutoAgent [8] compiles natural-language workflow descriptions into executable multi-agent pipelines. OWL [9] adds structured orchestration using a CAMEL-based Orchestrator–Worker pattern, where the orchestrator delegates subtasks to specialists via explicit transfer actions. For information-dense tasks, OpenAI Deep Research [1] combines web browsing, document parsing, and grounded synthesis to produce cited reports, while open Deep Research by Smolagents [10] reproduces the same workflow in the open-source domain using a Python-driven CodeAgent that reduces communication overhead. They jointly define today's design space for scalable generalist multi-agent solutions.

2.2 Domain-specific Agent

Domain-specific agents address the limitations of general-purpose LLMs in tasks requiring deep expertise. A primary approach to imbue such specificity is parametric adaptation through fine-tuning on domain-centric corpora, enabling models to learn specialized terminology, data patterns, and reasoning pathways. Notable examples include CRISPR-GPT [2], an agent designed to automate and enhance the design of CRISPR-based gene-editing experiments by leveraging domain knowledge and external tools, BrainGPT [11], a model clinically visual instruction-tuned for 3D brain CT radiology report generation , and DeepSeekMath [12], which demonstrates advanced mathematical reasoning after extensive training on mathematical texts. These methodologies are crucial for creating agents that are not only knowledgeable but also more reliable and effective within their specific operational contexts.

Beyond foundational knowledge embedding, the efficacy of domain-specific agents is significantly amplified by their ability to interact with and act upon their environment through tool integration. ChemCrow [3], for instance, leverages a suite of 18 expert-designed chemistry tools to perform complex tasks in organic synthesis, drug discovery, and materials design, showcasing enhanced performance in chemistry-related problem-solving. Other advanced approaches include Case-Based Reasoning (CBR), enabling agents such as DS-Agent [13] to learn from past expert solutions in fields like automated data science. WarAgent [14] exemplifies domain-specific simulation of macro-scale historical conflicts using LLM-based multi-agent systems, allowing country agents to reenact international relations dynamics and explore war triggers and outcomes. Similarly, BattleAgent [15] extends this line of work by incorporating multimodal inputs and fine-grained modeling of individual soldiers in historical combat, demonstrating how agent-based frameworks can reconstruct both strategic decisions and personal experiences in complex environments. These approaches collectively aim to produce agents that are not only competent in their domain but also robust and trustworthy in real-world settings.

2.3 Domain-specific Benchmarks

The evaluation of large language models (LLMs) is moving from broad assessments toward domain-specific benchmarks (DSBs). General-purpose tests often miss the detailed strengths and weaknesses of LLMs when applied in real-world settings. Common issues with standard benchmarks include limited coverage of field knowledge, poor alignment with practical tasks, vulnerability to data contamination that encourages memorization, and simple question formats that do not test multi-step or advanced reasoning [16, 17, 18, 19]. Questions about data quality and contamination have led to efforts such as HLE, which invests in carefully curated, contamination-resistant items [5]. As a result, the number of DSBs has grown rapidly. These new tests help guide model improvements, set realistic expectations for deployment, and pinpoint specific model capabilities in different areas [20, 21].

This trend towards specialization is evident across numerous domains. In software engineering, benchmarks like SWE-Bench [22] evaluate LLMs on resolving real-world GitHub issues, moving beyond simpler code generation tasks assessed by benchmarks like HumanEval or MBPP [23, 24, 22, 5]. The medical domain utilizes benchmarks such as PubMedQA [25] and the MultiMedQA suite [26] to assess medical knowledge and reasoning, although ongoing research seeks to improve alignment with clinical practice and address benchmark saturation [26, 17]. Legal AI evaluation employs benchmarks like LegalBench [20] for diverse legal reasoning tasks and specialized versions like LegalBench-RAG [21] focusing on retrieval-augmented generation crucial for fact-intensive legal work. For scientific and mathematical reasoning, SciBench [16] presents collegiate-level problems , RV-Bench [18] targets genuine mathematical understanding , and PHYBench [6] specifically assesses complex physics reasoning using problems from global exams and competitions. Finance has seen the development of benchmarks like FinanceQA [27], tailored to evaluate performance on tasks mirroring real-world financial analysis. Collectively, these domain-specific evaluations are crucial for probing deeper LLM capabilities and fostering the development of models suitable for specialized professional deployment.

2.4 History Benchmarks

In the domain of history, based on a subset of the Seshat Global History Databank, HiST-LLM evaluates the possession of expert historical knowledge of Seven Models. [28] However,there is a discrepancy between historical knowledge and historical research.HLE can be regarded as a combination of many domain-specific benchmarks. In terms of the historical questions in HLE, although the capabilities of LLMs in historical research is indeed evaluated (as opposed to the benchmarks mainly focused on historical knowledge), the total number of questions is only 56. [5] For existing benchmarks, there are limitations in both the scope of the questions and the disciplinary characteristics.

We need to detail several limitations in these benchmarks. The creators of Gaia [4] indicate that Gaia has three limitations—Missing evaluations: Shortage of evaluation of trace leading to the answer; On the cost of designing unambiguous questions: Shortage of ambiguous questions that may appear in the daily usage scenario; Lack of linguistic and cultural diversity: Shortage of questions in languages other than standard English.The creators of HiST-LLM also indicate the limitations of the database they use (the subset of the Seshat Gloval History Databank). First, their data are mostly from sources in English;Second, the expertise and background of the research assistants may influence the definition of variables. Third, the benchmark is only the reflection of the current recognition. In addition to these limitations, it is necessary to repeat that HiST-LLM is a benchmark designed for the evaluation of the possession of historical knowledge instead of the capabilities of historical research such as literature retrieval, historical source retrieval, historical analysis, historical source processing, and interdisciplinary.

Compared to the comprehensive benchmark like HLE, an independent domain-specific benchmark can be designed more suitable for a specific domain without the requirement of consistency and inclusiveness of different domains. To evaluate the capabilities of LLMs in the domain of historical research more adequately, creating a larger and more systemic dataset is necessary. If we take historical questions in HLE as an independent historical benchmark, there will be more space for improvement. First, the temporal and spatial scope of topics should be expanded with more questions. Second, the types and depth of functions to be evaluated should be richer. For example, we cannot find historical materials in the form of audios or video in historical questions of HLE. Third, in the context of the history discipline, the questions can be divided further in different groups according to their difficulty levels, thematic categories and evaluation criteria, which helps to constructing a more diverse but clear database.

Considering these limitations, the lack of independent domain-specific benchmarks suitable for historical research provides us with vast space for exploration.

3 HistBench Benchmark

3.1 Overview

HistBench is the first benchmark specifically designed to evaluate the capabilities of LLMs in historical reasoning. In the benchmark,two question types (exact match and multiple choice), six academic dimensions (e.g., source processing, interdisciplinary synthesis), and a three-tiered difficulty stratification based on task complexity are adopted. The benchmark contains **414 questions authored by more than 40 contributors**, including Ph.D. students and domain experts. All questions are grounded in real historical sources and methods and follow a standardized protocol that ensures academic rigor and traceability. Each question is annotated with metadata including answer explanation,data required,source material, targeted capability, and thematic category and is subjected to a three-stage review: preliminary screening (format and semantic check), LLM-based pre-evaluation and professional review (history of rigor and validity of reasoning) to ensure the quality of these questions in terms of both form and content,especially their academic value and challenges to current LLMs.

In terms of coverage, HistBench spans **29 ancient and modern languages, over 20 historical regions, and at least 36 subfields**. The questions are grounded in multimodal data, including printed texts, manuscripts, images, inscriptions, audio, and video. The benchmark also spans five major historical epoch, ranging from ancient history to the contemporary history, supporting diachronic analysis in diverse historiographic contexts. In all, these design criteria and dataset characteristics enable HistBench to serve as a robust diagnostic tool for evaluating LLMs in the humanities, with particular emphasis on capabilities of historical research like language mastery, historical materials processing and the understanding of historical analysis.

ID	Dimension	Description
1	Bibliographic Retrieval	The capability to locate information embedded in scholarly texts, monographs, or journal articles using digital or library-based search strategies.
2	Source Identification	The capability to recognize or locate specific historical sources, including manuscripts, digitized archives, or visual primary materials.
3	Source Processing	The capability to extract and interpret information from non-textual formats such as handwritten documents, historical images, audio, or video.
4	Historical Analysis	The capability to engage in historically grounded reasoning, including causal inference, ideological analysis, and interpretation of events or institutions.
5	Interdisciplinary Integration	The capability to draw upon methods and frameworks from adjacent disciplines (e.g., archaeology, linguistics, religious studies) to support historical understanding.
6	Cultural Contextualization	The capability to interpret cultural cues, metaphors, sentiment, and identity markers within historically situated discourse.

Table 1: Evaluation dimensions for historical reasoning tasks in HistBench.

3.2 Question Design and Structure

3.2.1 Question Types

HistBench, including two primary question types, is designed to evaluate both systematic historical research skills and in-depth historical reasoning:

Multiple-choice questions: These questions adopt a several-option format, where some of the options are designated as correct based on current scholarly consensus. The remaining distractors are carefully designed to be plausible yet incorrect, often reflecting common misconceptions or subtle historical variations. This type is suited to evaluate interpretive precision, historical knowledge, and the ability to differentiate between nuanced alternatives.

Exact match questions: These questions adopt a single precise answer, such as a date, name, location, or keyword. The LLMs needs to provide this precise answer. This type is suited to evaluate factual investigation, temporal reasoning, or targeted source analysis.

3.2.2 Question Format and Submission Protocol

All questions in HistBench are submitted using a standardized format developed by the project team to ensure scholarly rigor and traceability. Contributors are required to provide not only the question and answer, but also support metadata such as difficulty level, source materials, answer explanation, and bibliographic references. A detailed breakdown of the submission fields and formatting requirements is provided in Appendix A.1.

3.2.3 Evaluation Dimensions

As shown in Table1, Each question in the dataset is designed to evaluate specific dimensions of AI capabilities in historical research. These dimensions are modeled on the core challenges encountered by human researchers in historical research and embody key aspects of the research processes that characterize professional historical scholarship.

3.3 Difficulty Stratification

3.3.1 Difficulty Stratification Criteria

To support layered evaluation and targeted analysis, all questions in HistBench are categorized into three difficulty levels—Level 1 (Basic), Level 2 (Intermediate), and Level 3 (Challenging)—based on a structured rubric comprising six academic dimensions. Difficulty levels are determined based on six dimensions: (1) rarity of source knowledge, (2) linguistic complexity, (3) format heterogeneity, (4) perceptual accessibility, (5) interdisciplinary scope, and (6) reasoning complexity These dimensions inform both the formulation of questions and the assignment of difficulty levels.

This stratification provides a principled framework for evaluating how model performance differs with task complexity.

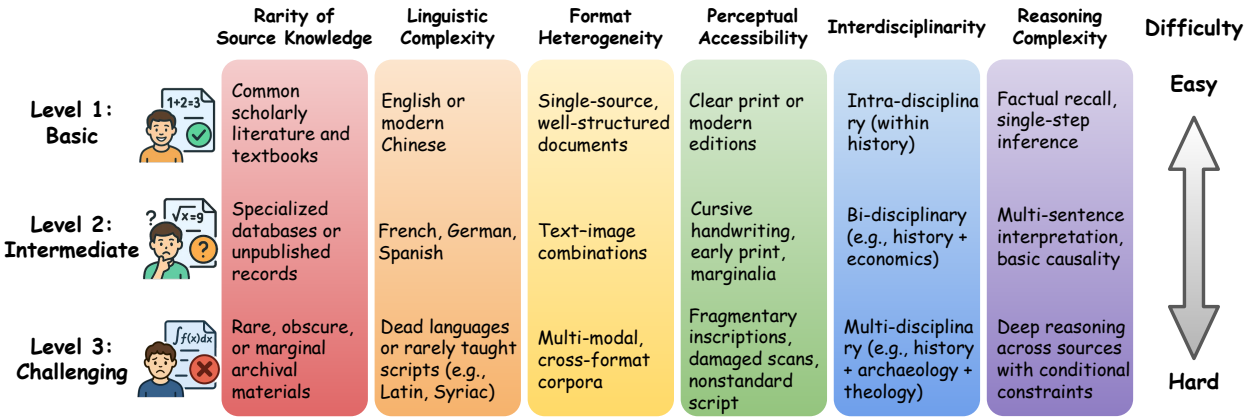


Figure 2: Difficulty level definitions across six structured evaluation dimensions.

3.3.2 Questions assignment in difficulty levels

By dividing contributors into different groups according to their different knowledge reserves, material processing capabilities, and reasoning capabilities, we assign contributors to formulate questions in different difficulty levels based on their groups. Our assignment plan is detailed as follows:

Level 1 (Basic): Authored by Research Assistants A total of 166 Level 1 questions are mainly formulated by research assistants with academic backgrounds in history. They are focused on extracting verifiable historical facts from historical materials or literature, interpreting well-documented primary sources to formulate questions. Tasks at this level emphasize factual investigation and basic material interpretation.

Level 2 (Intermediate): Authored by Graduate Researchers and Domain Experts The 172 Level 2 questions are mainly formulated by graduate students (MA and PhD level) and early-career researchers in history. Contributors are selected based on their specialization in particular subfields and are invited to design questions grounded in their areas of expertise. These tasks typically require source interpretation, basic causal or temporal research, and engagement with narrower but academically meaningful materials.

Level 3 (Challenging): Authored by Professors and Senior Scholars The 76 Level 3 questions are mainly contributed by senior scholars and university faculty with demonstrated experience in interpreting obscure sources, handling multi-lingual materials, and conducting interdisciplinary research. These tasks are designed to push the limits of LLM capabilities, drawing on rare texts, underdigitized archives, and cross-modal information integration.

In practice, we adjust difficulty levels of these questions according to both the inherent complexity of the tasks and a structured rubric. By involving contributors with varying levels of expertise—including research assistants, graduate students, and domain scholars—HistBench ensures that each question reflects a human-authored path of historical research appropriate to its level of difficulty. This multi-tiered design allows the benchmark to evaluate the capabilities of LLMs across a spectrum of historical tasks aligned with real-world scholarly practices. A selection of representative questions from each difficulty level is provided in Appendix A.2, offering concrete illustrations of the benchmark's structure and intended reasoning depth.

3.3.3 Question Review Protocol

All questions in HistBench undergo a three-stage review pipeline combining human oversight and automated evaluation, designed to ensure academic rigor, task clarity, and model discriminability. A question is only admitted into the final release if it passed all three stages described below.

Stage 1: Preliminary Screening (Format and Semantic Check)

The first round of review is conducted by the project team and trained research volunteers. This phase focuses on verifying structural completeness, semantic clarity, and question originality. Questions that are ambiguous, duplicated, or noncompliant with submission standards are flagged for revision, with detailed feedback sent back to the authors.

Stage 2: LLM-Based Pre-Evaluation

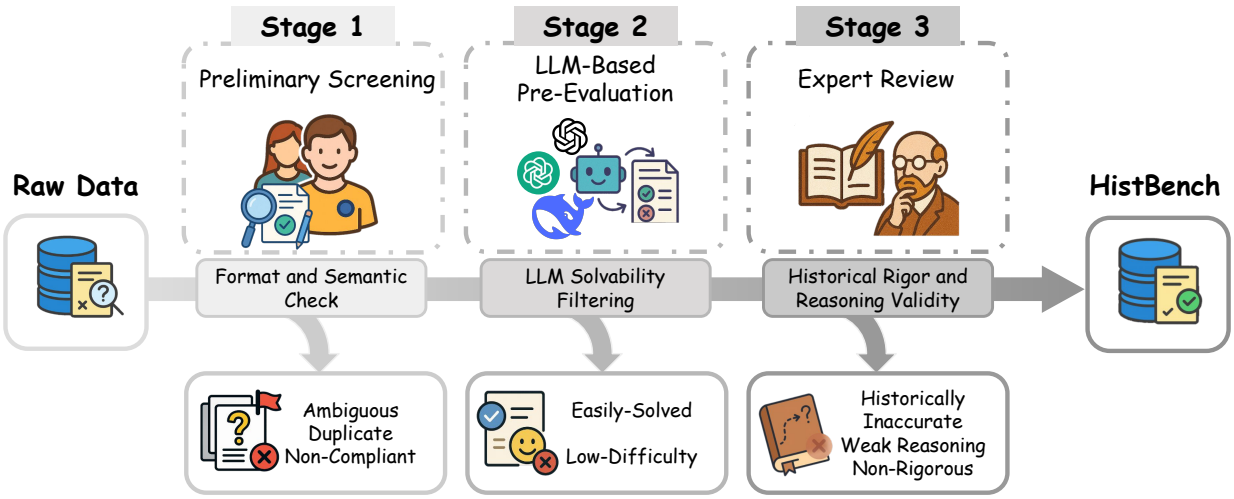


Figure 3: Multi-Stage Question Review Pipeline for HistBench

Before professional review, all questions are tested using multiple strong LLMs, including GPT-4o, GPT-o4mini, and DeepSeek-R1. Questions that are solved by more than two existing models without access to supporting materials are excluded from the dataset. This stage ensures that retained questions pose challenges beyond the capabilities of frontier general models.

Stage 3: Professional Review (Historical Rigor and Reasoning Validity)

Reviewers with academic backgrounds in history conduct the final round of review. The focus at this stage is on evaluating historical relevance of these questions, academic rigor, and reasoning quality. Reviewers assess whether the questions reflect valid historical methodologies, appropriate use of evidence, and internal logical coherence. Only questions deem to have academic value were accepted into the final benchmark.

3.4 Dataset Characteristics and Coverage

3.4.1 Quantitative Overview

HistBench contains a total of 414 questions, spanning diverse historical domains, periods, and geographic regions. The dataset includes two question formats: 306 exact match questions, which require precise, fact-based responses such as names, dates, or locations; and 108 multiple-choice questions, each consisting of one correct answer and three plausible distractors grounded in historical reasoning. The predominance of exact match tasks reflects an emphasis on fine-grained factual precision, while multiple-choice questions are designed to assess interpretive discrimination and the ability to reason across structured alternatives. This composition supports multi-faceted evaluation of both retrieval and inference capabilities in historical question answering.

3.4.2 Language Diversity

As shown in Figure 4, HistBench questions are grounded in materials written or originally produced in **29 modern and ancient languages**, capturing the multilingual nature of historical scholarship and the linguistic complexity of primary source analysis. The dataset includes widely used academic languages such as English, Chinese, Russian, Japanese, French, and German, as well as historically significant but less commonly used languages including Classical Chinese, Latin, Ancient Greek, and Tibetan.

This linguistic diversity presents substantive challenges for AI systems. It enables evaluation of multilingual retrieval, cross-lingual historical reasoning, and OCR robustness—addressing a core limitation of existing benchmarks that predominantly rely on English-language content. A complete list of languages and their frequencies is provided in Appendix A.3.

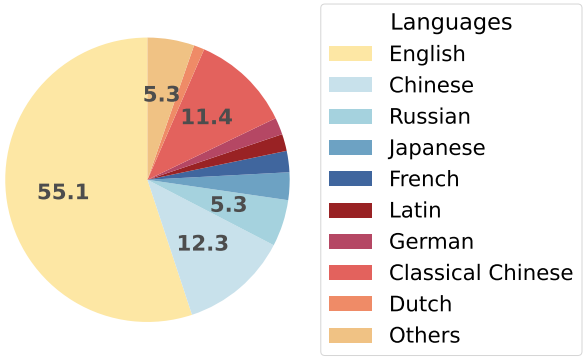


Figure 4: Language distribution in HistBench. The dataset covers 29 modern and ancient languages, including both widely used academic languages and historically significant lower-resource ones.

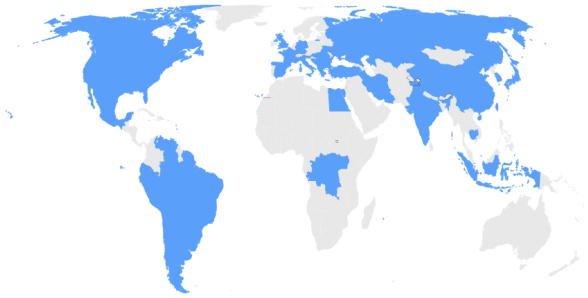


Figure 5: Geographic coverage of HistBench. Blue regions mark areas from which historical questions are drawn, spanning all inhabited continents and supporting both mainstream and underrepresented fields.

3.4.3 Data Source Modalities

The questions in HistBench are grounded in a diverse set of material sources, including printed texts, manuscript scans, visual materials, audio and video recordings, inscriptions, and multimodal combinations. Table 2 summarizes the number of questions associated with each type of source.

Material Type	Number of Questions
Visual materials (illustrations, photos)	96
Maps and schematics	18
Charts, diagrams, or tables	10
Manuscripts and handwritten sources	88
Audio recordings	9
Video content	5
Inscriptions or stone rubbings	14
Text-based questions (narrative excerpts)	64
Mixed text + image sources	10
Ancient or undeciphered scripts	22

Table 2: Distribution of source modalities in HistBench.

These sources are drawn from digitized archives, unpublished manuscripts, early print editions, and degraded or fragmentary artifacts. Many questions involve non-standard formats such as cursive handwriting, damaged inscriptions, or multilingual marginalia. Several questions combine multiple modalities, including text–image pairings or audiovisual content.

Crucially, many of these materials remain inaccessible to existing OCR tools and frontier LLMs, either due to visual complexity, symbolic ambiguity, or contextual dependence. By incorporating such cases, HistBench evaluates not only recognition and parsing but also a model's ability to reconstruct historical meaning from incomplete or opaque data.

3.4.4 Domain and Geographic Coverage

HistBench includes coverage of at least **36 historical subdomains**, **over 20 national or civilizational areas**, and multiple cross-cultural comparative themes, which ensures that HistBench covers a large temporal and spatial scope of topics. At the domain level, the dataset encompasses:

- **Political, social, and cultural history**, including diplomatic history, gender studies, intellectual history, and identity politics;
- **Classics and ancient civilizations**, covering Greco-Roman studies, philology, epigraphy, and early textual traditions;

- **Art and visual culture**, including art history, iconography, visual semiotics, and historical image interpretation;
- **Material culture and archaeology**, including material artifact studies, heritage reconstruction, and excavation-based historiography;
- **Environmental and climate history**, such as the historical study of climate shifts, ecological regimes, and resource management;
- **History of science and medicine**, including early scientific institutions, cross-cultural scientific exchange, botany, astronomy, and traditional medicine;
- **Economic and institutional history**, such as labor systems, taxation, urban planning, bureaucratic organization, and legal codification;
- **Interdisciplinary and comparative studies**, including global history, translation history, mythology, and civilizational entanglements.

Geographically, HistBench draws from nearly all inhabited continents, including East Asia, Europe, North America, Latin America, the Middle East, and Africa, as shown in Fig. 5. It includes questions grounded in region-specific expertise (e.g., Dunhuang studies, Slavic paleography) as well as underrepresented areas such as papyrology and Siberian ethnography, ensuring broad civilizational and epistemic representation.

In addition to thematic and regional diversity, HistBench evaluates the capabilities of historical research across the full temporal arc of human history. To capture both general historical literacy and specialized period knowledge, we adopt a five-part periodization commonly used in Western and global historiography: Prehistory, Ancient History, the Middle Ages, Modern History, and Contemporary History.[29] [30]This framework reflects canonical divisions in comparative history and supports systematic coverage of distinct epistemic contexts. Table 3 summarizes the distribution of questions across historical periods.

Historical Period	**Number of Questions**
Ancient History (to ∼500 CE)	90
Medieval History (500–1500)	85
Early Modern History (1500–1800)	95
Modern History (1800–1945)	80
Contemporary History (1945–present)	64

Table 3: Chronological coverage of questions in HistBench.

4 HistAgent

HistAgent is a complex agent system designed to enhance LLM-based historical reasoning by integrating specialized agents and external tools. Unlike basic retrieval systems, it supports academic search, multimodal input (e.g., manuscripts, images, audio), and outputs fully cited responses grounded in both primary and secondary sources (see Fig. 6).

HistAgent consists of two core components. A central Manager Agent orchestrates the pipeline: it decomposes queries, dispatches subtasks, verifies evidence, and assembles the final cited response. Surrounding it is a set of specialized agents, each handling a specific task such as web search, literature retrieval, OCR, translation, image or audio analysis.

Among them, the Literature Search Agent is key for scholarly retrieval and citation. It follows a multi-stage protocol focused on academic databases and bibliographic validation. We detail its design and workflow in the following section.

4.1 Agent Architecture and Roles

4.1.1 Manager Agent

The Manager Agent parses the user request to identify required modalities, and then runs a CodeAct-style loop as a Smolagents `CodeAgent`. In each iteration, the LLM emits a Python code snippet that calls sub-agent functions or tools; the snippet executes in a secure sandbox and writes its output to shared memory. The Agent then validates each result (checking exact quotes and bibliographic data) and repeats until task completion criteria are met. In the final step, the agent synthesizes all validated outputs into a structured, citation-complete response.

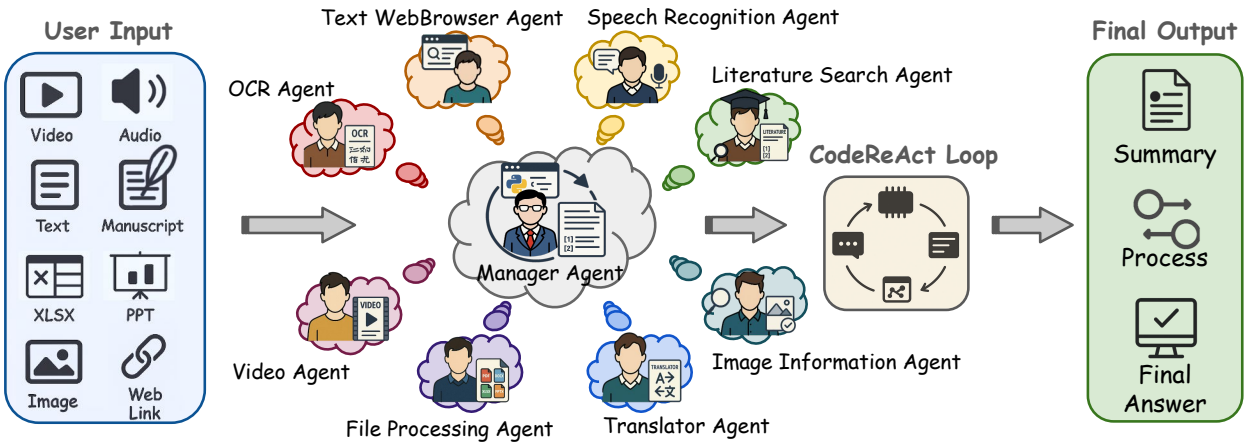


Figure 6: The architecture of HistAgent, an agent for historical reasoning. The system takes multimodal user inputs (e.g., text, audio, video, images, documents) and coordinates specialized agents through a centralized Manager Agent. Each agent is responsible for dedicated tasks such as OCR, speech recognition, literature search, translation, and file processing. The Manager Agent orchestrates these agents within a CodeAct loop to iteratively refine reasoning and evidence validation, ultimately generating summarized, processed, and fully cited academic answers.

Agent	Modality	Core Tools
Text WebBrowser	Text	Multi-step web search and page parsing (`LocalGoogleSearchTool`, `VisitTool`).
Image Information	Image	Reverse image search and provenance analysis (`GoogleLensSearchTool` with SSIM filtering).
Literature Search	Scholarly text	Peer-review retrieval and PDF parsing (Scholar websites, `SpringerDownloadAndParseTool`).
File Processing	Documents	Typed file parsing: `PDFTool`, `DOCXTool`, `XLSXTool`, `PPTXTool`.
OCR	Image text	Manuscript transcription (Transkribus, Asian-script OCR).
Speech Recognition	Audio	Whisper-based transcription with LLM correction.
Translator	Multilingual text	Bidirectional translation with provenance preservation.
Video	Video	Frame extraction (yt-dlp, OpenCV).

Table 4: Function Overview of HistAgent Specialist Agents.

4.1.2 Specialist Agents

An overview of all agents and their primary toolkits is presented in Table 4.

Manager Agent The Manager Agent is the central coordinator. It parses user requests, selects which agent to invoke, gathers their outputs, checks completeness, and synthesizes a final response.

Functionality. The Manager Agent parses the user request to identify required modalities, then runs a ReAct-style loop as a Smolagents `CodeAgent`. In each iteration, the LLM emits a Python code snippet that calls sub-agent functions or tools; the snippet executes in a secure sandbox and writes its output to shared memory. The Agent then validates each result (checking exact quotes and bibliographic data) and repeats until a `final_answer(...)` call. At that point, it merges all verified outputs into a single, citation-ready response.

Text WebBrowser Agent The Text WebBrowser Agent handles open-domain web searches and browsing. It simulates a multi-step search and extracts structured content from web pages using a suite of navigation and inspection tools.

Functionality. The agent starts by refining the user's query via `LocalGoogleSearchTool` into an optimized search prompt. It can also perform standard queries via `SearchInformationTool`. For each target, it invokes `VisitTool` to load the page, `DownloadTool` to save binary files when needed, and `ArchiveSearchTool` to retrieve historical

snapshots. When encountering long pages, it scrolls with `PageUpTool` and `PageDownTool`, and it locates specific terms using `FinderTool` and `FindNextTool`. Non-HTML content, like plain text files, PDFs, or video transcripts, is converted to text through `TextInspectorTool`. All outputs are recorded to shared memory for downstream integration.

Image Information Agent The Image Information Agent focuses on visual inputs. It performs reverse image search and follows up with targeted page visits to uncover context and provenance of the visual input.

Functionality The Image Information Agent is selectively invoked when the task involves image content that may offer contextual or evidential value. Upon receiving an image, the Image Information Agent uploads it to a public host and runs a reverse search using `GoogleLensSearchTool`. The system extracts associated links, titles, and descriptions from matched pages to identify how the image is used online, e.g., in auction listings, academic articles, or museum databases. To improve match quality, the agent optionally computes similarity scores (e.g., SSIM) between the original image and search results to highlight high-confidence matches. To gather in-depth metadata, it visits selected pages with `VisitTool_Image`. By grounding image interpretation in its real-world usage context, the agent enables accurate and verifiable historical reasoning over visual inputs.

Literature Search Agent The Literature Search Agent is a specialized module for retrieving and grounding answers in peer-reviewed academic sources. It combines an LLM-driven browser workflow with multiple tool wrappers, covering Google Books, Google Scholar, general web searches, and Springer Nature's API, to locate exact phrases, download accessible PDFs, and extract verbatim quotes along with full citation metadata.

Functionality. The agent issues a multi-stage search: it first queries Google Books (with random domain rotation and robots.txt compliance) via `BookMatchExtractorTool` and `DirectGoogleBooksCrawlerTool`, extracting highlighted snippets and page numbers; if needed, it proceeds to Google Scholar using `LiteratureSearchingTool` and `RelevantLiteratureFinderTool` for broader article discovery; it falls back on `GeneralBrowserTool` for additional context. Freely accessible PDFs are downloaded and parsed with `SpringerDownloadAndParseTool` (via LlamaParse) or `SpringerSearchTool`/`SpringerStructuredSearchTool` for structured Springer Nature queries. Throughout, it preserves exact wording, records source URLs, citation counts, publication details, and assembles all findings into a coherent, verifiable response. it can locate sources that support factual claims, provide historical context, or contain exact wording required for exactMatch tasks. It returns full bibliographic metadata, quoted excerpts, and links to original publications, thereby ensuring academic-level faithfulness and verifiability.

File Processing Agent This agent manages non-HTML files—documents, spreadsheets, presentations, and images—by routing them to the appropriate tool.

Functionality. When a file is received, the file processing Agent automatically detects its type and selects the corresponding tool: `PDFTool` for PDFs, `DOCXTool` for Word documents, `XLSXTool` for spreadsheets, and `PPTXTool` for presentations. For images that require analysis beyond OCR, it uses `ImageAnalysisTool` to extract charts or figures. All extracted text or structured data is returned in a format that downstream agents or the Manager Agent can incorporate into their reasoning.

OCR Agent The OCR Agent specializes in extracting textual information from images using optical character recognition. It is invoked for screenshots, scanned documents, historical manuscripts, or photos containing embedded text. The agent supports multiple languages and detects the language in the image. Upon invocation, it returns the raw text content detected, enabling HistAgent to convert unstructured visual inputs into machine-readable text for further processing.

Functionality. When given an image path, the agent loads and encodes the file and uses an LLM to determine whether the content is best handled by a specialized OCR model for Asian scripts or by a Transkribus-based OCR service for Western-language manuscripts. For Western texts, it publishes the image to a public URL, submits it to the Transkribus engine, waits for a PAGE XML transcription, and extracts the detected lines. For Asian scripts, it sends the image data directly to the dedicated OCR model. The resulting raw transcription is then passed through an LLM prompt that repairs recognition errors and preserves historical or stylistic features. Both the original and refined transcripts are saved to a `.txt` file and returned. If no valid text is extracted, the agent falls back to generating a detailed visual description via the LLM, highlighting any readable text, symbols, or key visual elements.

Speech Recognition Agent The Speech Recognition Agent converts audio files (MP3, WAV, etc.) into text using Whisper for transcription and an LLM for error correction, summary, and key-point extraction, enabling HistAgent to incorporate oral historical sources.

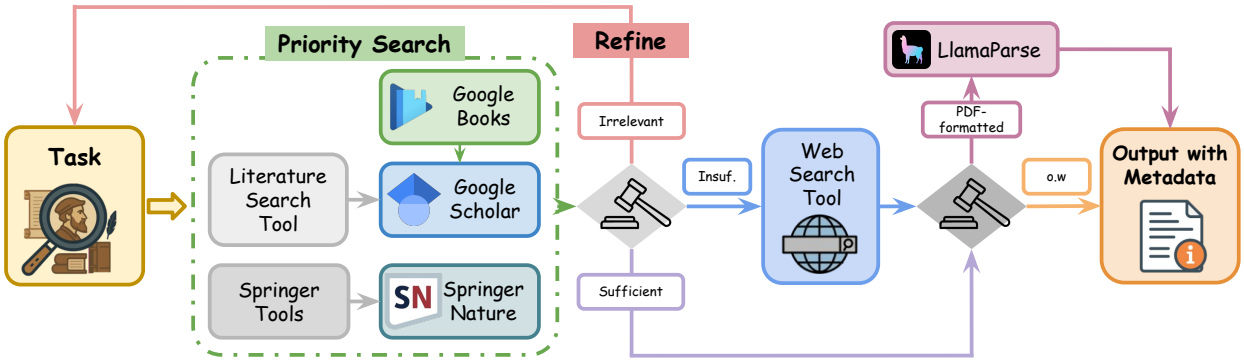


Figure 7: Overview of Literature Search Agent. It is specifically designed for academic searching, including a priority search module and web search tools.

Functionality. When given an audio file path, the agent verifies its existence and measures its size; if the file exceeds 25 MB, it divides the recording into equal segments; otherwise, it uses the file as a whole. Each segment is sent to Whisper for transcription, and the concatenated raw transcript is submitted to an LLM prompt that preserves all content while correcting recognition errors and generating an "Optimized Transcription" section, a brief "Summary," and a "Key Points" list. Both original and refined texts are saved to a `.txt` file in the output directory for humans' reference, and a formatted string containing both versions is returned, with any errors caught and reported.

Translator Agent The Translator Agent converts text into a specified target language, including support for both widely spoken and less common languages like Armenian and Sanskrit, and delivers a clear, formatted output showing the original and translated text.

Functionality. The Translator Agent handles automatic translation of textual content between multiple languages. It automatically detects the source language and supports both widely used and lower-resource languages. In historical tasks, it is particularly useful for translating foreign-language sources such as Armenian manuscripts, Sanskrit inscriptions, or early regional documents in Latin or Classical Arabic. The translated output allows HistAgent to reason across linguistic boundaries and integrate multilingual content into its historical analysis pipeline.

Video Agent The Video Agent downloads the video from the given link and extracts still frames at a user-specified rate to support visual analysis.

Functionality. When given a video URL, the agent downloads the best-quality video with yt-dlp, uses OpenCV to extract frames at the specified rate, saving each as a timestamped JPEG, and writes a summary file containing the video's title, duration, resolution, frame count, and output directory. It then returns a brief report with those key details and file locations.

4.2 Literature Search Agent

The Literature Search Agent is a highly specialized component within the HistAgent framework, engineered as a ReAct-based agent driven by a large language model (LLM). Its core architectural design focuses on emulating a meticulous academic research process through a structured, multi-stage interaction with a curated set of web-based tools, primarily implemented through browser-use [31], which can connect the agent to the real Chrome browser instance. This setup allows the agent to operate within the user's active browser profile, inheriting authentication states and personalized settings for seamless access to academic resources.

Its internal architecture is centered around a protocol-driven retrieval engine that ensures the reliability, interpretability, and academic integrity of the information it gathers. The agent **prioritizes scholarly sources** such as Google Scholar and Springer Nature, employs a **modular toolkit** for interfacing with both academic databases and general-purpose web sources, and allows users to configure search depth and behavior via **interpretable parameters**. Retrieved content is enriched with bibliographic metadata, quoted evidence, and **full traceability** links to support verification and integration into downstream workflows.

4.2.1 Priority Search Protocol

A core challenge in automating academic research is ensuring that retrieved information aligns with scholarly standards. To address this, our system introduces a priority-based retrieval mechanism that favors academically reputable sources over general-purpose content. This design choice reflects the distinct requirements of academic tasks, where the reliability of information is critical. The protocol acts as a persistent control signal for the agent, influencing both the decision to invoke tools and the interpretation of retrieved results.

4.2.2 Toolkits for Literature Search

Efficient execution of the academic-first retrieval strategy requires structured tool use. To this end, our system integrates a suite of specialized retrieval tools, each targeting different tiers of source quality. All source queries, are first conducted via the `LiteratureSearchingTool` to search from scholarly databases, and then to `GeneralBrowserTool` interface for general search. The queries will also use more specialized APIs when needed. For example, Springer-specific queries are handled using the `SpringerSearchTool`, and focused access to Google Books is enabled through the `DirectGoogleBooksCrawlerTool`. Most of the retrieved documents can be extracted directly, while pdfs could be parsed via the `Llama-Parse` API for structured processing. This modular architecture ensures that each retrieval action is consistent with the source prioritization strategy established earlier.

4.2.3 Human-controllable Agent Customization

Autonomous retrieval must adapt to varying task demands and resource limitations. Our system provides users with explicit control over agent behavior through a small set of interpretable parameters. These include limits on reasoning steps and re-planning intervals, which allow users to trade off between search depth and computational cost. Such configurability is critical for real-world deployment, where users may need to enforce practical constraints while preserving the academic quality of the output.

4.2.4 Clearly labeled Sources

Trust and traceability are essential in academic applications. To meet these needs, our system ensures that all retrieved content is accompanied by clearly labeled source metadata. This includes bibliographic information, retrieval URLs, and quoted evidence, all linked to the reasoning trace. These annotations allow users and downstream systems to verify claims, reproduce retrieval steps, and integrate results into formal academic workflows.

5 Experiment

5.1 Experiment Setup and Datasets

We evaluate HistAgent on three benchmarks covering challenging historical and real-world general tasks. Also, we have extensive baselines.

HistBench. HistBench is a history-specific benchmark we construct, containing 414 history questions categorized into three difficulty levels (Level 1–3).

GAIA. To test the generalization of our framework, we also test our HistAgent on the validation set of GAIA benchmark[4], which includes 165 tasks.

HLE History Subset. We consider the HLE history subset of 56 questions from the Humanity's Last Exam benchmark[5].

We compare our HistAgent against open Deep Research (ODR-smolagents)[10], an open-source reproduction of OpenAI's Deep Research agent by HuggingFace/smolagents. It uses a Manager agent and a Text Web Browser Agent with a visualizer tool and a file processing tool. Both systems are run under identical environments for fairness. We use the same underlying language model (GPT-4o) for both HistAgent and ODR-smolagents, ensuring that differences in performance stem from the agent architecture rather than the base model. Each question is posed to the agent in a fresh session. We limit the total number of tool invocations (agent calls) per question to a fixed budget to prevent infinite loops and to enforce a fair comparison. Both systems output a final answer for each question, which we compare to the ground-truth answer. Additionally, we compare our results with DeepSeek-R1 [32], GPT-4o [33], o4-mini-high [34], o3 [34], and Grok 3 [35] on HistBench for a thorough comparison. All these LLMs are equipped with online search capabilities, enabling them to be very strong baselines.

5.2 Evaluation Metrics for Different Benchmarks

We use accuracy as the primary evaluation metric, defined as the percentage of questions answered correctly by the agents or LLMs.

Metric for HistBench. For the HistBench benchmark, we measure accuracy as the percentage of the 414 questions for which the agent's response is judged correct by both the LLM judge and professtional validation. The evaluation proceeds in three steps:

- **Structured Output:** Each response is composed of a concise final answer and a structured reasoning summary (shown in Appendix E.3) that logs tools used, information sources, and step-by-step logic.
- **LLM Judging:** We run LLM as a judge to issue a binary judgment ("Correct"/"Incorrect") based on semantic equivalence, completeness of key facts, and logical coherence. The prompt template is an adaptation of the evaluation prompt of HLE, which is shown in Appendix E.2.
- **Human Expert Validation:** To validate the LLM's judgment quality, we randomly sample 100 examples across Level 1, Level 2, and Level 3 tasks. Each sample includes both the final answer and its reasoning summary. These are reviewed by the original authors of the questions, who assess whether the correct answer could have been obtained merely by coincidence, verify the reliability and factual accuracy of every cited source, and confirm that the sequence of reasoning steps genuinely supports the final answer. An answer is labeled "Correct" only if it satisfies all of these criteria; otherwise, it is marked "Incorrect". This human expert evaluation not only provides an external check on model-assigned labels but also reveals edge cases and ambiguities in question phrasing or evaluation criteria, allowing us to iteratively refine and improve the overall quality and consistency of our HistBench.

We define accuracy on HistBench as the percentage of the 414 questions for which the response is both "Correct" by the LLM judge and "Correct" by the author (100 samples selected). We report this metric overall and separately for Level 1, Level 2, and Level 3.

Metric for GAIA. We employ the official scoring function published on Hugging Face. Each GAIA question has a standard answer key; the scoring function handles normalization (e.g., case, punctuation) and computes an exact-match score. We report the overall GAIA accuracy as defined by that function.

Metric for HLE History Subset. For the 56 expert-level questions in the HLE History Subset (curated from Humanity's Last Exam), evaluation mirrors the initial stages of the main HLE benchmark protocol. Each response must first provide a Structured Output, comprising a concise final answer and a detailed reasoning summary (as described for the HLE benchmark before). Subsequently, we employ LLM-based Judging, where an LLM assesses the response for semantic equivalence to the ground truth, factual completeness, and logical coherence, issuing a binary judgment ("Correct"/"Incorrect") as the HLE official scoring function released on github. Accuracy for this subset is then calculated as the percentage of questions deemed "Correct" by the LLM judge.

5.3 Results and Analysis

5.3.1 Performance on HistBench

Table 5 shows that HistAgent, based on GPT-4o, achieves 36.47% pass@2 on HistBench, outperforming all baselines, including those with web search.

Compared to the open-source agent ODR-smolagents (also using GPT-4o), HistAgent improves by 11.35 points (a 45.2% relative gain), demonstrating the impact of its enhanced tools, search mechanisms, and file handling. Against GPT-4o with web search, HistAgent improves by 17.87 points, indicating that its performance stems not just from tool access, but from better tool selection, sequencing, and integration.

Interestingly, while ODR-smolagents slightly outperforms GPT-4o with web search (25.12% pass@2 vs. 18.60%), gains are limited, emphasizing that domain-specific agent is essential. HistAgent meets this need through components like handwriting OCR, multimodal analysis, translation, and a scholar search agent with query optimization.

Despite using a weaker base model, HistAgent also matches or outperforms stronger closed-source models like o3 and o4-mini-high. This highlights the value of its reasoning structure and tool design. Inference cost further makes HistAgent suitable for cost-sensitive scenarios where models like o3 are impractical.

Future work will explore applying HistAgent to stronger foundation models (e.g., o3, o4-mini-high) to assess the upper bounds enabled by its architecture.

Agent/Model		Level 1	Level 2	Level 3	Average
HistAgent(gpt-4o)	pass@1	28.92	23.84	32.89	27.54
	pass@2	36.14	35.47	39.47	36.47
ODR-smolagents(gpt-4o)	pass@1	16.27	23.26	22.37	20.29
	pass@2	20.48	28.49	27.63	25.12
Deepseek-R1:online		11.45	18.60	11.84	14.49
GPT-4o:online		13.86	21.51	22.37	18.60
o4-mini-high:online		28.92	31.98	31.58	30.68
o3:online		18.07	41.28	43.42	32.37
Grok 3:online		13.25	19.77	22.37	17.63

Note: All standalone large language models are used **with web search** capabilities. The best value in each column is highlighted with darker blue, and the second-best score with lighter blue.

Table 5: Performance accuracy (%) on HistBench

5.3.2 Performance on the HLE History Subset

Focusing on these 56 expert-level history questions in HLE History Subset, we observe that HistAgent markedly outperforms both baselines. HistAgent achieves a Pass@1 accuracy of 28.6% (16 out of 56 questions), whereas the ODR-smolagents baseline reaches 17.9% (10 out of 56 questions), and the GPT-4o baseline achieves 8.9% (5 out of 56 questions). This subset is exceptionally difficult—indeed, even state-of-the-art models have been reported to achieve less than 20% accuracy on comparable expert-level history questions—so all systems understandably struggle on many of these prompts. However, HistAgent answers 60% more questions correctly at Pass@1 than the ODR-smolagents baseline and more than triples the performance of the direct GPT-4o baseline. More details are provided in Table 6.

System	Pass@1	Pass@2	Pass@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

Table 6: Performance on the HLE History Subset (56 Questions, LLM Judged, %). To ensure fair comparison, both HistAgent and ODR-smolagents are based on GPT-4o.

5.3.3 Performance on GAIA Validation

We also test HistAgent on GAIA benchmark, which achieves 60.00% pass@1 accuracy, showing that our domain-specific adaptations don't hinder HistAgent's competitive performance on real-world general tasks.

We evaluate our HistAgent on the 165-question validation subset, using the accuracy as the main metric. The split includes 53 Level 1, 86 Level 2, and 26 Level 3 questions, covering open-book fact finding, tool use, and multimodal reasoning. Our HistAgent, based on the GPT-4o, answers 99 questions correctly and reaches an overall accuracy of 60.00% pass@1. The baseline, open Deep Research by HuggingFace/smolagents, records 55.15% on the same split. These results show that HistAgent, as a domain-specific agent, can generalise reliably beyond its original scope. More details are provided in Table 7, which presents the complete level-wise breakdown.

Agent	Model	Average	Level 1	Level 2	Level 3
HistAgent	Claude 3.7 Sonnet	**60.00**	69.81	61.63	34.62
ODR-smolagents	o1	55.15	67.92	53.49	34.62

Table 7: Performance accuracy (%) on the GAIA validation set (pass@1)

6 Conclusion

In summary, we introduce HistBench, which is a benchmark for evaluating historical reasoning in AI, and HistAgent, which is a specialized AI agent that outperforms LLMs and other agents on historical tasks by using domain-specific tools and workflows.

A HistBench Benchmark

A.1 Submission Template Format

To illustrate how these elements are applied in practice, Table 8 provides two sample entries from different difficulty levels.

- **(a) Difficulty Level:** Assigned as Level 1, 2, or 3 based on rubric criteria (see Section 4.3).
- **(b) Question Prompt:** A clear and concise question targeting a specific historical issue requiring domain expertise.
- **(c) Required Data:** Source materials referenced or used (e.g., documents, images, audio/video).
- **(d) Answer:** A definitive, validated response—either as a selected option or short text span.
- **(e) Answer Explanation:** A concise justification based on evidence and reasoning.
- **(f) Source References:** URLs or bibliographic citations supporting the answer.
- **(g) Topic/Methodology:** Thematic or methodological classification (e.g., diplomatic history, material culture).
- **(h) Contributor Name:** Full name of the author.
- **(i) Contributor Affiliation:** Institutional affiliation at the time of contribution.

This format ensured each question adhered to standards of academic transparency and could be independently reviewed and validated.

Field	**Q001 (Level 1)**	**Q002 (Level 3)**
ID	Q001	Q002
Answer Type	Multiple Choice	Exact Match
Question	What year did the Treaty of Westphalia end the Thirty Years' War?	Translate and date the following Latin inscription found on a Roman milestone in Gaul.
Data Requirements	Modern European history textbook excerpt	Image of milestone inscription (Latin)
Answer	1648	Circa 220 CE
Answer Explanation	The Treaty of Westphalia was signed in 1648, ending the Thirty Years' War in Europe.	The inscription, typical of the Severan dynasty period, was dated to around 220 CE using epigraphic style.
Source Materials	*Merriman, A History of Modern Europe*, p. 203	*Corpus Inscriptionum Latinarum*, Vol. XIII
Thematic Category	Political History	Epigraphy / Classical Studies
Evaluation Criteria	Factual recall	Source processing; temporal inference; Latin translation
Contributor's Name	XX	XXX
Contributor's Affiliation	XXX	XXX

Table 8: Vertical Format Sample Entries from HistBench

A.2 Sample Questions Across Difficulty Levels

To illustrate the scope and structure of questions in HistBench, we present three annotated examples—one for each difficulty level. These samples highlight variation in required source processing, historical reasoning, and interdisciplinary complexity.

- **Level 1 (Basic)** – Source verification and factual retrieval
- **Level 2 (Intermediate)** – Text-image synthesis and temporal reasoning
- **Level 3 (Challenging)** – Multilingual, multimodal, and interdisciplinary integration

Level_1: Retrieval of Rare Data

Question:

Which ancient greek papyrus is this excerpt below taken from:

κληδὼν δὲ κλίτα ὡς Πίτωνα τοξεύσας,

τῶι Ζεὺς δαιδουχεῖ παρ' ἡμέραν κατ' ἄγαν,

τῶι γᾶς ἐν βώλοις ξανθοὶ τίκτονται καρποί.

A.Papyrus of Vienna G29825

B.Berlin Papyrus 6870

C.Papyrus Oxyrhynchus 2436

D.Papyrus Oslo 1

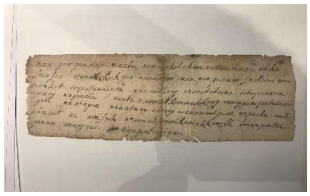
Level_2: Multilingual Analysis

Question:

Один профессор однажды сказал в интервью: «Только очень наивный инопланетянин мог бы подумать о землянах как о единой цивилизации. На самом деле он обнаружил бы по меньшей мере девять цивилизаций, в которых есть своя структура.» В частности, он отметил: «Евразийская цивилизация (Россия, Украина, Белоруссия)... Отличается своеобразием социальной психологии народов, образом жизни в труде, быте, досуге. Здесь сложилась особая алкогольная цивилизация с вековыми питейными традициями, невиданными ни на Западе, ни на Востоке. И особая, неслыханная нигде ругань. Невиданная нигде смесь самоотверженности и скандальности, западной культуры и восточного быта. Словом, Азия, но в европейском обличье.» Какой он профессор?

Level_3: Augmented Comprehensive Ability

Question:



According to this deed from early eighteenth-century New England, what is the relationship between Johan Pattukqua and the sachem Quequenap (or Quequenab), who conveyed the land?

A. Johan Pattukqua was a witness at the meeting in which the deed was issued.
B. Johan Pattukqua was the scribe of the sachem.
C. Johan Pattukqua was a previous owner of part of the land.
D. Johan Pattukqua was the receiver of the land E. Johan Pattukqua was a friend of the sachem.

Level_2: Better Multimodal Capability

Question:

At 04:42 - 08:59, the screen shows a public green space. Which of the following historical facts about the place is incorrect?

A. There is a monument in the center of this saqure-like public space. This monument got its name from the pro-Union rallies held there on the eve of the Civil War.

B. When this clip was filmed, there was no underground garage in the area.

C. Based on the layout and facilities of the square in this clip, we can infer that the video was shot after the end of the Spanish-American War.

D. This place is on the list of California Historical Landmarks (CHL).

Level_3: Debris Identification

Question:

In all, how many vehicles and men were captured by Duo You（多友） in the attack on the Xian Yun（玁狁） tribe?

Level_3: High-level Paleography

Question:

širirnalandaram-ka eltd[ilär] anta tägdökdä kamag [kuvrag] yıgılmıš ärdi, samt[so ačari] olar-nı birlä körüšü t[ükättök] dä š[ilaba]ḍre ačari bašınt[a][yegir]mi [a]čari-ka samtso ača]rig uduzturup [drmaguptake] ačari-ka yüküntür [gäli]… …dı-lar , d(a)rmaguptake [ačari] ärsär , kayken lüši [tegmä] šilabadre ačari käntü öz

Who is the samtso ačari mentioned in this Old Uighur text?

A. Šilabaḍre B. Faxiang C. Xuanzang D. Šingqu Säli E. Aśoka

A.3 Language Distribution in HistBench

Table 9 provides the full list of languages represented in the HistBench dataset, along with their frequencies. These include both modern languages and historical scripts, reflecting the multilingual nature of historical research tasks. This diversity supports the evaluation of cross-lingual capabilities in AI systems, including translation, OCR, and historical reasoning across languages.

Language	Count
English	228
Chinese	52
Russian	22
Japanese	13
French	10
Latin	8
German	8
Classical Chinese	47
Dutch	5
Tibetan	2
Armenian	2
Arabic	2
Khitan	2
Ancient Greek	2
Khmer	1
Indonesian	1
Old Tibetan	1
Sanskrit	1
Old Uyghur	1
Middle Polish	1
Aramaic	1
Danish	1
Bosnian	1
Italian	1
Macedonian	1
Yukaghir	1

Table 9: Languages represented in HistBench questions.

B Limitations

Our HistBench has two limitations. First, all the questions are closed questions with an only exact answer to simplify the evaluation criteria, which lead to a monotony in the types of questions. In fact, handling of questions with an open answer is a critical part of historical research. The evaluation of questions with an open answer needs an intricate and dynamic evaluation criterion. The determination of this evaluation criterion is difficult and requires further exploration. Second, although contributors are required to formulate questions with objective answers, the design of questions is still inevitably influenced by the contributors' cognition and the current research limitations. For example, due to differences in material selection and interpretation,the academic world may have different answers to an objective question, while the contributors may choose the information they have accessed and believe to provide the answers for the questions. This appears to be an inevitable drawback.

C Broader Impacts

D Declaration of LLM Usage

This research involves the use of large language models (LLMs) as core components of the experimental framework. Specifically, the HistAgent (our method) and ODR-smolagents (baseline) both rely on GPT-4o as the underlying language model. The purpose of our experiments is to isolate and compare the performance impact of different agent

architectures, with the LLM held constant across systems to ensure a fair comparison. The differences in outcome are therefore attributed to the design of the agent frameworks, not the underlying model capabilities.

Each agent operates in a fresh session per question. We impose a fixed budget on the number of tool invocations (agent calls) to avoid infinite loops and normalize the computational cost across runs. The output of each agent is a final answer that we compare against ground-truth responses.

In addition to GPT-4o-based systems, we also benchmark against DeepSeek-R1, GPT-4o, o4-mini-high, o3, and Grok 3 on HistBench. These models are used in their native agentic or search-augmented setups, consistent with how they are deployed in real-world scenarios. All these models have online search capabilities and demonstrate strong reasoning abilities in open-domain tasks.

The LLMs are strictly part of the technical comparison and evaluation of agent performance, which forms the core contribution of the work.

E Details in Experiment

E.1 Experiment Compute Resources

M1 chips, 512 MB, 20 minutes for 1 question.

E.2 Prompt Template

JUDGE_PROMPT Template

```
JUDGE_PROMPT = """You are a fair evaluator.  Judge whether the following [response]
to [question] is semantically consistent with the [correct_answer] below.
[question]:  {question}
[response]:  {response}
[correct_answer]:  {correct_answer}
When you judge, consider only whether the core meaning and all necessary key points
in the response match the correct answer.  Even if wording or format differs, treat
equivalent semantics as correct.  Treat missing key points or any substantive error
or omission as incorrect.  For numerical answers, a small rounding difference is
acceptable.  Tolerate substantive deviations from the correct answer.  If the
extracted_final_answer is a more specific instance of the correct_answer (for
example, ‘‘Pieter Schenk II’’ vs ‘‘Pieter Schenk’’), and it still contains the core
string of the correct_answer, treat it as correct.
Please output exactly in the format and criteria specified below:
extracted_final_answer:  The final exact answer extracted from the [response].  Put
the extracted answer as ’None’ if there is no exact, final answer to extract from
the response.
reasoning:  Explain why the extracted_final_answer is correct or incorrect based
on [correct_answer], focusing only on if there are meaningful differences between
[correct_answer] and the extracted_final_answer.  Do not comment on any background
to the problem, do not attempt to solve the problem, do not argue for any answer
different than [correct_answer], focus only on whether the answers match.
correct:  Answer ’yes’ if extracted_final_answer matches the [correct_answer] given
above, or is within a small margin of error for numerical problems.  Answer ’no’
otherwise, i.e.  if there is any inconsistency, ambiguity, non-equivalency, or if
the extracted answer is incorrect.
confidence:  The extracted confidence score between 0% and 100% from [response].
Put 100 if there is no confidence score available."""
```

E.3 Summary Case Study

Summary Case Study: Summary for Level 2_30

Question ID: level_2_30
Question: What is the name of the man in this picture?
Answer Type: exactMatch
Data Requirement: level_2_30.png
Data Type: file
Our Answer: Veit Rudolph Speckle
Correct Answer: Veit Rudolph Speckle / Veit Rudolf Specklin
Is Correct: Yes
File: Historical/Historical/level_2_30.png
Model: gpt-4o

Summary Report
1. Tools Used and How They Were Used:
- **Image Analysis and Reverse Image Search**:
- Purpose: To identify the individual in the image "Historical/Historical/level_2_30.png".
- Method: Used the `image_information_agent` to perform a reverse image search to identify the name attached to the image and gather related historical context.
- Result: The reverse search linked the image to Veit Rudolph Speckle, associated with botanical illustrations for Leonhart Fuchs' "De Historia Stirpium".
- **Literature Search Agent**:
- Purpose: To find scholarly literature verifying the exact match for Veit Rudolph Speckle's association with Fuchs' work.
- Method: Used the query "Veit Rudolph Speckle is one of the engravers associated with Leonhart Fuchs' 'De Historia Stirpium.'".
- Result: Confirmed findings that Speckle was a renowned engraver for this botanical work, highlighted as responsible for key woodcut illustrations.
- **Web Search**:
- Purpose: To access additional context and verify the scholarly data.
- Method: Conducted a search on Google and accessed Google Books, Google Scholar, and other scholarly sources.
- Result: Successfully retrieved confirmation and additional scholarly and book references regarding the engravers and contributors to Fuchs' herbal.
2. Detailed Information Sources:
- **Wikidata**
- URL: Wikidata: Veit Rudolph Speckle
- Quote: "The primary identification linked it to Veit Rudolph Speckle, as noted on Wikidata..."
- Credibility: Open user-contributed database but verifies common scholarly facts.
- **Article in "The World of Plants in Renaissance Tuscany" by Cristina Bellorini**:
- Quote: "Veit Rudolph Speckle was an engraver for Leonhart Fuchs' 'De Historia Stirpium'."
- Credibility: Academic book providing historical insight into botanical studies during the Renaissance.
- **Google Scholar Sources**:
- Quotes: "Veit Rudolph Speckle was responsible for the woodcut engravings in Leonhart Fuchs' herbal." and "Recognized as 'by far the best engraver in Strasbourg...".
- URLs: Retrieved through exploratory searches within Google Scholar.
- Credibility: Google's academic resource is known for aggregating reputable, peer-reviewed materials.
3. Reasoning Process and Logic Steps:
- **Identification**:
- Initial determination of the man in the image was made through reverse image search, identifying him as Veit Rudolph Speckle.
- **Verification**:
- Utilized literature searches on academic databases (Google Scholar, books) to verify Speckle's role in creating illustrations for Fuchs' botanical book.

- **Cross-verification**:
- Multiple sources including books and academic papers were consulted to confirm Speckle's work as an engraver for Leonhart Fuchs' book "De Historia Stirpium."
- **Exclusion of Other Answers**:
- The reverse image search did not present any alternative credible identity, leading to a focused inquiry on Speckle which was consistently supported by scholarly resources.

4. Answer Quality and Reliability Analysis:

- **Reliability**: High
- Given the cross-corroboration from reliable academic texts and reputable databases (Google Scholar, academic books), the reliability is high.
- **Assumptions, Weaknesses, Uncertainties**:
- Assumptions largely relied on the historical accuracy maintained by sources. Lack of web search limits alternative verifications.
- **Sufficiency and Consistency**:
- The evidence gathered was sufficient, consistent, and convergent from independent, reliable sources, affirming the credibility of the information.
- **Suggestions for Improvement**:
- Include a broader web search to capture any contemporary assessments or potential misattributions regarding this artwork or engraver. Suggested keywords: "Veit Rudolph Speckle", "Leonhart Fuchs botanical engravings".
- **Web Search Observation**:
- Although literature and specific academic sources provided strong backing, the integration of broader web search could improve the reliability by encompassing wider perspectives or additional public domain resources.

References

[1] OpenAI. Introducing deep research.

[2] Kaixuan Huang, Yuanhao Qu, Henry Cousins, William A Johnson, Di Yin, Mihir Shah, Denny Zhou, Russ Altman, Mengdi Wang, and Le Cong. Crispr-gpt: An llm agent for automated design of gene-editing experiments. *arXiv preprint arXiv:2404.18021*, 2024.

[3] Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

[4] Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*, 2023.

[5] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

[6] Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Haoxu Zhang, Yi Hu, et al. Phybench: Holistic evaluation of physical perception and reasoning in large language models. *arXiv preprint arXiv:2504.16074*, 2025.

[7] Xun Jiang, Feng Li, Han Zhao, Jiaying Wang, Jun Shao, Shihao Xu, Shu Zhang, Weiling Chen, Xavier Tang, Yize Chen, et al. Long term memory: The foundation of ai self-evolution. *arXiv preprint arXiv:2410.15665*, 2024.

[8] Jiabin Tang, Tianyu Fan, and Chao Huang. Autoagent: A fully-automated and zero-code framework for llm agents. *arXiv e-prints*, pages arXiv–2502, 2025.

[9] Mengkang Hu, Yuhang Zhou, Wendong Fan, Yuzhou Nie, Bowei Xia, Tao Sun, Ziyu Ye, Zhaoxuan Jin, Yingru Li, Zeyu Zhang, Yifeng Wang, Qianshuo Ye, Ping Luo, and Guohao Li. Owl: Optimized workforce learning for general multi-agent assistance in real-world task automation, 2025.

[10] Aymeric Roucher, Albert Villanova del Moral, Thomas Wolf, Leandro von Werra, and Erik Kaunismäki. 'smolagents': a smol library to build great agentic systems. `https://github.com/huggingface/smolagents`, 2025.

[11] Cheng-Yi Li, Kao-Jung Chang, Cheng-Fu Yang, Hsin-Yu Wu, Wenting Chen, Hritik Bansal, Ling Chen, Yi-Ping Yang, Yu-Chun Chen, Shih-Pin Chen, et al. Towards a holistic framework for multimodal llm in 3d brain ct radiology report generation. *Nature Communications*, 16(1):2258, 2025.

[12] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

[13] Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. Ds-agent: Automated data science by empowering large language models with case-based reasoning, 2024.

[14] Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *arXiv preprint arXiv:2311.17227*, 2023.

[15] Shuhang Lin, Wenyue Hua, Lingyao Li, Che-Jui Chang, Lizhou Fan, Jianchao Ji, Hang Hua, Mingyu Jin, Jiebo Luo, and Yongfeng Zhang. Battleagent: Multi-modal dynamic emulation on historical battles to complement historical analysis. *arXiv preprint arXiv:2404.15532*, 2024.

[16] Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023.

[17] Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. Medical large language model benchmarks should prioritize construct validity. *arXiv preprint arXiv:2503.10694*, 2025.

[18] Zijin Hong, Hao Wu, Su Dong, Junnan Dong, Yilin Xiao, Yujing Zhang, Zhu Wang, Feiran Huang, Linyi Li, Hongxia Yang, et al. Benchmarking large language models via random variables. *arXiv preprint arXiv:2501.11790*, 2025.

[19] Jan-Micha Bodensohn, Ulf Brackmann, Liane Vogel, Anupam Sanghi, and Carsten Binnig. Unveiling challenges for llms in enterprise data engineering. *arXiv preprint arXiv:2504.10950*, 2025.

[20] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279, 2023.

[21] Nicholas Pipitone and Ghita Houir Alami. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*, 2024.

[22] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

[23] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[24] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[25] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

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

[27] Spencer Mateega, Carlos Georgescu, and Danny Tang. Financeqa: A benchmark for evaluating financial analysis capabilities of large language models. *arXiv preprint arXiv:2501.18062*, 2025.

[28] Jakob Hauser, Daniel Kondor, Jenny Reddish, Majid Benam, Enrico Cioni, Federica Villa, James Bennett, Daniel Hoyer, Pieter Francois, Peter Turchin, et al. Large language models' expert-level global history knowledge benchmark (hist-llm). *Advances in Neural Information Processing Systems*, 37:32336–32369, 2024.

[29] Daniel Woolf. *A Global History of History*. Cambridge University Press, 2011.

[30] John H. Arnold. *History: A Very Short Introduction*. Oxford University Press, 2000.

[31] Magnus Müller and Gregor Žunič. Browser use: Enable ai to control your browser, 2024.

[32] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[33] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[34] OpenAI. Introducing openai o3 and o4-mini, 2025. Accessed 15 May 2025.

[35] xAI. Grok 3 beta — the age of reasoning agents, 2025. Accessed 15 May 2025.