

Extending AI for Research to the Humanities: A Multi-Agent Framework for Evidence-Grounded Scholarship

Yating Pan^{1,2} Jiajun Zhang^{1,2} Jun Wang^{1,2,*} Qi Su^{3,2,†}

¹Department of Information Management, Peking University

²Research Center for Digital Humanities, Peking University

³School of Foreign Languages, Peking University

ytpan25@stu.pku.edu.cn 2601111285@stu.pku.edu.cn

junwang@pku.edu.cn sukia@pku.edu.cn

Abstract

LLM-based research agents have advanced rapidly in science and engineering, where research is organized around executable experiments, code, and quantitative signals. Humanities scholarship, however, requires interpretive, evidence-grounded arguments over primary sources, whose value rests on faithful quotation, verifiable provenance, and deep interpretation. Existing research agents use general planning, tool use, and reflection, leaving these scholarly operations implicit. To address this gap, we introduce **SPIRE** (Scholarly-Primitives-Inspired Research Engine), a multi-agent framework that realizes Scholarly Primitives, a typology of basic humanities scholarship practices, as cooperating roles over a multi-scale close-reading substrate of passages, intra-context graph communities, and cross-context semantic clusters. On a benchmark of peer-reviewed papers in classical Chinese and Greco-Roman Latin scholarship, **SPIRE** recovers substantially more cited primary-source evidence and receives higher blind human and LLM ratings on four scholarly quality dimensions than LLM, RAG, and generic agentic baselines. Retrieval-tier and agent ablations, with retrieval-volume and answer-length controls, trace its advantage to targeted retrieval, structured evidence selection, and claim-evidence binding. Code and benchmark data are released at <https://github.com/YatingPan/SPIRE>.

1 Introduction

LLM-based agents have recently achieved strong performance in science and engineering. In fields such as chemistry and materials science, they design experiments, analyze data, and test hypotheses

(Boiko et al., 2023; Ghafarollahi and Buehler, 2024; Chen et al., 2025). In computer science and software engineering, they survey literature, write and test code, and plan experiments (Lu et al., 2024; Romera-Paredes et al., 2024; Schmidgall et al., 2025; He et al., 2025; Jimenez et al., 2024; Zhang et al., 2025; Zhuge et al., 2024). They also conduct quantitative analysis and simulate participants or populations in the social sciences (Park et al., 2023; Horton et al., 2023; Argyle et al., 2023; Li et al., 2023; Piao et al., 2025; Gao et al., 2024). Advances in workflow orchestration (Wu et al., 2023), multi-agent collaboration (Chen et al., 2024b), retrieval-augmented generation (Lewis et al., 2020; Gao et al., 2023), reasoning and acting loops (Yao et al., 2023), self-reflection (Shinn et al., 2023), and reinforcement learning (Guo et al., 2025) enable these systems to coordinate complex research tasks. Across these domains, research is typically organized around executable procedures, verifiable numerical outcomes, code, or data analysis.

This mode of agentic assistance remains less developed in text-centered humanities scholarship, whose core practices require different capabilities.¹ Humanities scholarship is qualitative, interpretive, and evidence-grounded. Scholarly training emphasizes selecting and closely reading primary sources, mining and weighing evidence for and against a claim, and developing original interpretations. Each interpretation rests on faithful quotation, verifiable provenance, and primary-source evidence (Smith, 2016). Compared with fixed experimental procedures or reproducible code and computation, humanistic inquiry places greater weight on evi-

*Corresponding author.

†Co-corresponding author.

¹By humanities research, we focus on text-centered scholarship in history, philosophy, literature, philology, and so on, excluding computationally oriented areas such as quantitative history and digital humanities (Moretti, 2013; Jockers, 2013; Underwood, 2019; Rockwell and Sinclair, 2016).

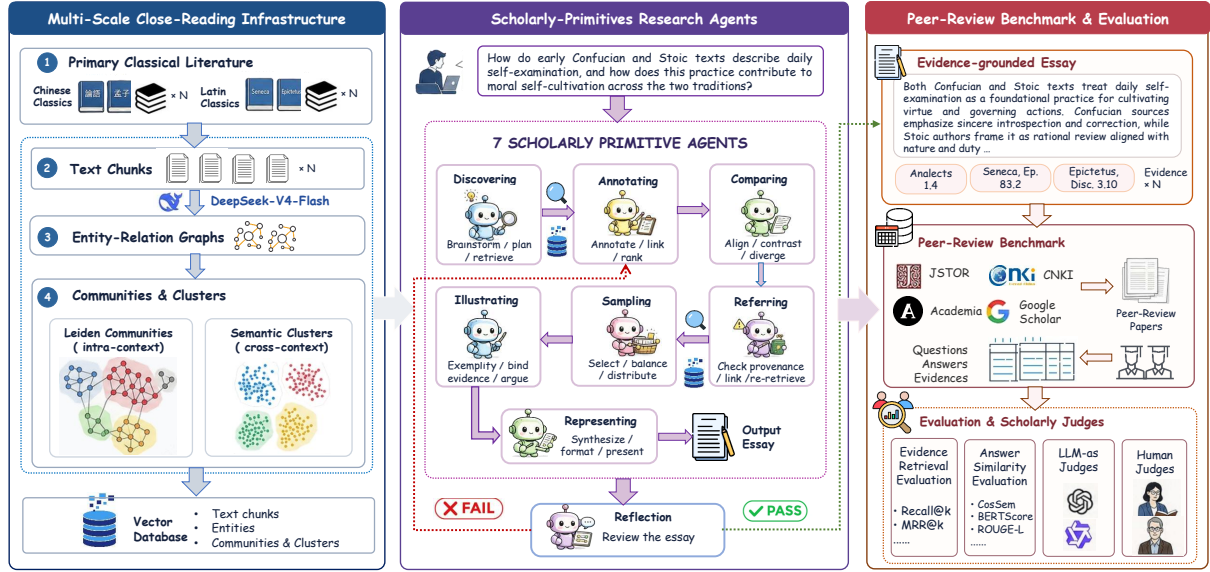


Figure 1: Overview of **SPIRE**: compiles classics into a multi-scale close-reading store, answers a research question with seven scholarly-primitive agents over an EvidencePool, and evaluates on a peer-reviewed-paper benchmark.

dential sufficiency and interpretive depth. Multiple well-supported readings may remain possible, leaving outcomes open-ended and difficult to verify with a single quantitative signal. Agentic assistance must therefore make scholarly operations and evidence paths explicit.

Although humanities scholarship is interpretive and open-ended, scholarly training has long been organized around recurring practices that offer an explicit basis for agentic workflows. Scholarly Primitives theory characterizes the basic activities shared across humanities scholarship as discovering, annotating, comparing, referring, sampling, illustrating, and representing (Unsworth, 2000). Building on this perspective, we operationalize these practices as an executable multi-agent workflow. **SPIRE** (Scholarly-Primitives-Inspired Research Engine) coordinates agents over a typed primary-source EvidencePool to discover sources, annotate evidence, compare perspectives, verify provenance, sample passages, bind citations, and synthesize a citation-grounded argument. To our knowledge, **SPIRE** is among the first multi-agent research frameworks designed explicitly for evidence-grounded interpretive scholarship and models a complete path from source discovery to argumentative synthesis.

The EvidencePool in turn requires a retrieval substrate that reflects how humanists engage with primary texts. Close reading grounds interpretation and argument in the detailed selection and examination of primary sources. RAG and its variants

improve semantic access by embedding passages and, increasingly, graphs or structured knowledge (Lewis et al., 2020; Gao et al., 2023; Edge et al., 2025; Luo et al., 2025; Han et al., 2025). Humanistic reading also situates passages across multiple contextual scales, including explicit comparison with related works (Kristeva, 1980) and recurring conceptual formations across texts (Lovejoy, 1936). We therefore build a multi-scale close-reading substrate for the scholarly-primitive agents: passage embeddings for direct text evidence, entity-relation graph communities for intra-context relational evidence, and semantic clusters for cross-context recurrent evidence.

Evaluating such assistance requires measures that reflect both evidence grounding and interpretive essay quality. To this end, we construct a peer-reviewed-paper benchmark over classical Chinese and Greco-Roman Latin scholarship, extracting each paper’s research question, main findings, and cited primary-source evidence. Given only the title and research question, a system must perform the scholarly steps and produce an evidence-grounded essay. We evaluate two complementary dimensions: cited-evidence recovery at the sentence, section, and work levels, testing whether the system retrieves primary sources used in the reference paper; and blind scholarly evaluation by both expert human reviewers and LLM judges assessing answer accuracy, argument depth, coverage, and evidence quality. Compared with Naive LLM, Text RAG, GraphRAG, and generic agen-

tic baselines, **SPIRE** attains 44.3% evidence recall versus 30.0% for the strongest baseline and receives higher judge scores on all four aspects. Dynamic and frozen-pool ablations separate iterative evidence acquisition from post-retrieval curation. Retrieval-volume and answer-length controls test the advantage under fixed volume and comparable response length.

Our contributions are as follows:

1. We introduce **SPIRE**, a multi-agent framework for evidence-grounded humanities scholarship that operationalizes Scholarly Primitives as coordinated agent workflows for interpretive research over primary sources.
2. We propose a multi-scale close-reading substrate combining passages, graph communities, and semantic clusters to support direct evidence grounding, contextual interpretation, and recurrent conceptual analysis.
3. We construct a peer-reviewed-paper benchmark of 406 papers centered on primary-source evidence and scholarly judging, providing a reusable testbed for future humanities research agents.

2 Background and Related Work

2.1 LLM agents for research and their evaluation

LLM agents have emerged as a major paradigm for AI-assisted research (Zheng et al., 2025), especially in domains with explicit operational or verification signals. In the natural sciences, agents connect LLMs with laboratory APIs and domain-specific tools to plan and execute experiments (Boiko et al., 2023; Bran et al., 2024). In computer science, they automate parts of the research cycle, from ideation to code execution and paper writing (Lu et al., 2024; Li et al., 2024; Schmidgall et al., 2025). In quantitative social science, they simulate individuals, populations, and economic activity (Park et al., 2023; Li et al., 2023; Piao et al., 2025). Agent performance is commonly evaluated through task completion and executable outcomes. For example, ScienceAgentBench evaluates generated programs and execution results on data-driven scientific tasks (Chen et al., 2025), while SWE-bench tests whether agents resolve software issues through repository tests (Jimenez et al., 2024).

Research agents, however, remain less common in the humanities. Humanistic inquiry centers on

evidence-based interpretation of primary-source texts and may support multiple defensible interpretations. This limits the fit of agents and benchmarks built around executable tasks or reproducible quantitative results. **SPIRE** addresses the gap by modeling established scholarly practices as an agent workflow and evaluating primary-source evidence retrieval and essay quality against domain-specific criteria.

2.2 Retrieval-augmented generation for research

In addition to research agents, retrieval-augmented generation (RAG) supports AI-assisted research through literature reading, evidence synthesis, and knowledge retrieval. It grounds model outputs in retrieved passages to reduce unsupported claims and provide citation-backed answers (Lewis et al., 2020). PaperQA retrieves and assesses passages across full-text scientific articles (Lálá et al., 2023), while OpenScholar searches millions of papers and synthesizes citation-backed responses through retrieval and self-feedback (Asai et al., 2024). Graph-based variants extract structured relational knowledge from documents to strengthen global retrieval. GraphRAG uses entity relations and community summaries, while HyperGraphRAG represents higher-order relations through hyperedges (Edge et al., 2025; Luo et al., 2025). Such systems have supported knowledge-base question answering in medical and technical domains (Wu et al., 2025; Kim et al., 2026; Knollmeyer et al., 2025).

In the humanities, RAG has mainly supported access to source collections. FolkRAG retrieves cultural heritage materials, while Topic-RAG uses topic-guided retrieval over historical newspapers (Kelly et al., 2025; Murugaraj et al., 2025). These systems improve information access. Direct research assistance must also organize and interpret textual evidence across multiple contexts. **SPIRE** addresses this need through a multi-scale close-reading substrate of passages, graph communities, and semantic clusters.

2.3 Scholarly Primitives and close-reading theory

Humanities research paradigms provide the domain structure needed to adapt the general agents and retrieval methods above. Experimental science is often organized into repeatable, fine-grained procedures. Humanistic inquiry follows more flexible paths, yet its long-established practices can still be

summarized as shared scholarly operations. Scholarly Primitives identifies these operations as discovering, annotating, comparing, referring, sampling, illustrating, and representing (Unsworth, 2000). Work on scholarly information practices and humanities visualization similarly describes humanistic research through information seeking, interpretation, organization, and representation (Palmer et al., 2009; Drucker, 2011). We use this framework because it turns broad scholarly practices into explicit roles that agents can perform.

Agent roles define the research workflow. Primary-source interpretation also requires a retrieval substrate that situates evidence within textual, historical, and conceptual contexts. Close-reading traditions offer two complementary perspectives. Contextual reading, associated with Skinner and the Cambridge School, links meaning to historically attested texts, authors, intentions, and arguments (Skinner, 1969). Recurrent reading, drawing on Lovejoy’s history of ideas, traces conceptual formations across periods, works, and traditions (Lovejoy, 1936). It can surface relevant evidence when no direct textual link is documented. Together, these modes organize evidence within and across contexts. **SPiRE** maps them to multi-scale retrieval: passages for direct quotation, graph communities for intra-context relational evidence, and semantic clusters for cross-context recurrent evidence.

3 Methods: SPiRE

3.1 Overview

Building on the scholarly primitives and close-reading modes of §2.3, **SPiRE** combines a multi-scale evidence store with seven cooperating research agents. A user supplies a classical-text corpus, an entity/relation schema, and a research question. The system returns an evidence-grounded essay with traceable primary-source citations (Figure 1). Let $\mathcal{C}$ be a corpus segmented into text units u , and let $\mathcal{Q}$ be the research question. **SPiRE** produces a ranked EvidencePool $\mathcal{E}(\mathcal{Q})$ and an essay $A(\mathcal{Q})$ grounded in that pool.

The system has two connected components. The multi-scale close-reading infrastructure (§3.3) organizes $\mathcal{C}$ as passages, an entity-relation graph, intra-context communities, and cross-context clusters. The scholarly-primitive workflow (§3.2) turns a question into sub-questions, retrieves evidence from these levels, and passes a shared Evidence-

Pool through seven agents. The agents interpret, compare, verify, and select evidence before synthesizing the essay. Reflection reviews the result and can request targeted evidence for up to two rounds.

3.2 Scholarly-primitive agents and the EvidencePool

SPiRE realizes the seven Scholarly Primitives (§2.3) as cooperating agent roles. They share an evolving EvidencePool that connects retrieval, evidence analysis, and writing. Table 1 summarizes their operations, while Figure 2 follows one complete run.

3.2.1 Agent roles and the shared EvidencePool

Primitive	Main operation	Effect on pool or output
$\mathcal{D}$ Discovering	Plan sub-questions and search three tiers.	Initialize the EvidencePool.
$\mathcal{N}$ Annotating	Interpret and score retrieved passages.	re- Annotate and re-rank evidence.
$\mathcal{P}$ Comparing	Compare question-defined source groups.	Add parallels and divergences.
$\mathcal{R}$ Referring	Check provenance and source links.	Validate and expand the pool.
$\mathcal{S}$ Sampling	Balance question, work, and tradition coverage.	Select the evidence subset.
$\mathcal{I}$ Illustrating	Exemplify claims with retrieved passages.	Bind claims to evidence IDs.
$\mathcal{T}$ Representing	Organize evidence-bound claims.	Produce the final essay.

Table 1: The seven scholarly primitives and their effects on the shared EvidencePool and research output.

Definition 1 (EvidencePool). The *EvidencePool* $\mathcal{E}_k = [e_1, \dots, e_n]$ is the shared research workspace after workflow stage k . Each item $e_j = (u_j, m_j, r_j, z_j)$ contains a primary-source text unit u_j , source metadata m_j , retrieval provenance r_j , and agent-added information z_j . The metadata identifies its author, work, chapter or section, language, and canonical locator. Retrieval provenance records the sub-question and retrieval tier that produced the item. Agent-added information can include an interpretation, relevance score, comparison relation, provenance check, or citation binding.

Discovering initializes $\mathcal{E}_0$. Later agents can add evidence, attach interpretations, revise its ranking, organize comparative relations, verify sources, and select a final subset. The primary text and its locator remain linked throughout the workflow. Agents can therefore quote a passage directly and read it within its section and work context.

3.2.2 Query planning and evidence acquisition

Given $\mathcal{Q}$, Discovering ($\mathcal{D}$) produces a query plan $\pi(\mathcal{Q}) = (\{q_i\}, \text{lang}, c, \text{anchors}, \text{hints})$. The plan identifies source languages, literal work and concept anchors, and the comparison scope c . It also decomposes the question into two to four sub-questions q_i , each tagged by granularity and analytical perspective. Retrieval hints supply related concepts, native terms, and candidate works that broaden the search.

Each q_i queries three tools over the multi-scale store (§3.3). The passage tier L retrieves local text units and their entities. The community tier G retrieves intra-context reports and expands them to supporting units. The cluster tier I retrieves cross-context semantic affinities and follows them to their supporting entities and passages. Their ranked union initializes the EvidencePool:

$$\mathcal{E}_0(\mathcal{Q}) = \text{Rank} \left(\bigcup_i \bigcup_{t \in \{L, G, I\}} T_t(q_i) \right). \quad (1)$$

The ranking combines explicit work and anchor matches with semantic similarity. Every result retains its source locator, retrieval tier, and originating sub-question. Broad community and cluster matches can therefore be traced back to citable primary-source passages.

3.2.3 Orchestration, synthesis, and reflection

After Discovering creates $\mathcal{E}_0$, the orchestrator runs

$$\mathcal{D} \rightarrow \mathcal{N} \rightarrow \mathcal{P} \rightarrow \mathcal{R} \rightarrow \mathcal{S} \rightarrow \mathcal{I} \rightarrow \mathcal{T} \rightarrow \text{Reflection}.$$

Annotating relates each passage to a sub-question and assigns interpretive relevance. Comparing organizes the evidence groups specified by the query plan, such as named works, authors, or traditions. It records parallels, divergences, unilateral concepts, and inverted emphases. Referring checks work and section provenance, follows explicit source connections, and retrieves evidence for missing named sources. Sampling then selects a balanced subset across sub-questions, works, and traditions.

Illustrating forms evidence-grounded claims and binds them to stable evidence IDs. Representing organizes these claims into the final essay. A separate Reflection step reviews question coverage, evidence balance, identifiable gaps, and source support. A diagnosed gap produces a targeted retrieval query and returns the new evidence to annotation

and the downstream agents. The process ends when the review passes or reaches two rounds.

Figure 2 illustrates the sequence for a comparison between Cicero’s *De Re Publica* and the *Analects*. These two works define the comparison groups. Discovering creates three sub-questions and retrieves from all three tiers. Sampling retains eight Chinese and six Latin evidence items, and Illustrating binds four arguments to these fourteen items before Representing writes the essay. Per-agent prompt contracts and output schemas appear in App. F.

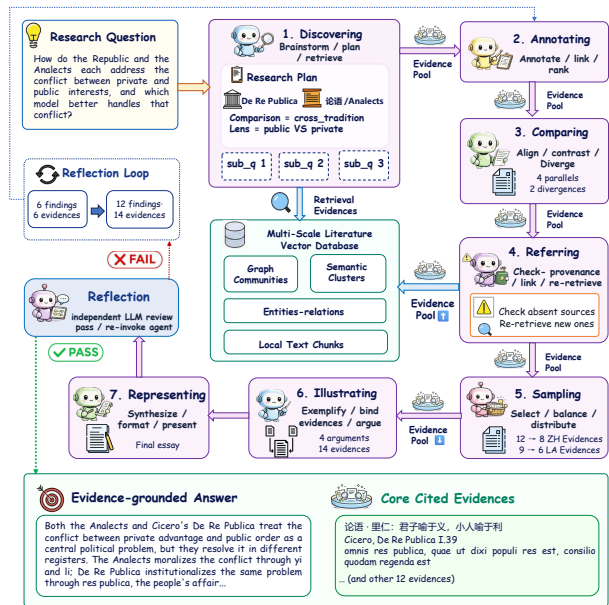


Figure 2: A cross-tradition run on Cicero’s *De Re Publica* and the *Analects*. Discovering forms the query plan and initial EvidencePool. Later agents annotate, compare, validate, select, and synthesize its evidence; Reflection can trigger targeted retrieval.

3.3 Multi-scale close-reading infrastructure

3.3.1 Corpora and text units

We instantiate **SPIRE** on two classical corpora central to humanities scholarship: classical Chinese (*Zhongguo Xueshu Mingzhu Tiya* (Zhou, 1992)) and Greco-Roman Latin (*The Latin Library* (The Latin Library, 2026)). History graduate students annotate every work with metadata including author, date, and alternative titles. Both span antiquity to the modern period and are standard primary sources in history, philology, and literary study. Table 2 reports their sizes and time spans.

We segment each corpus with `tiktoken` into 1000-token text units with 100-token overlap, giving $U = \{u_1, \dots, u_M\}$. The unit size preserves

	Classical Chinese	Classical Latin	Total
Documents	226	710	936
Text Chunks	51,726	35,117	86,843
Entities	133,686	75,318	209,004
Relation edges	286,123	176,405	462,528
Leiden communities	7,452	4,647	12,099
HDBSCAN clusters	9,795	6,480	16,275
Time span	8c BCE–19c CE	5c BCE–20c CE	–

Table 2: Classical sources corpus and retrieval substrate

the surrounding argument, speaker, and citation context needed for classical passages. Overlap reduces fragmentation at unit boundaries. Each unit remains linked to its work and section metadata.

3.3.2 Building the multi-scale store

A user defines an entity/relation schema, and an LLM extracts a typed graph from the text units. Our exemplar is *conceptual history*, a task shared across history, literary studies, and philology. The schema contains three entity types, *Concept*, *Person*, and *Work*, together with free, text-grounded relations. Its few-shot prompt contains one Chinese and one Latin example (App. F.2). DEEPSEEK-V4-FLASH performs extraction. Format validation, deduplication, and cross-unit entity merging yield a typed graph $\mathcal{G} = (\mathcal{V}, \mathcal{R})$, where $\mathcal{R} \subseteq \mathcal{V} \times \mathcal{V}$ contains directed semantic relations. Per-tradition counts appear in Table 2, and schemas for other fields appear in App. E.

Two higher levels organize the graph at different contextual scales. The **intra-context** level captures corpus-attested relations among works, persons, and concepts, following contextual reading (Skinner, 1969). Leiden community detection (Traag et al., 2019) groups these connected entities. The **cross-context** level captures recurrent conceptual affinity among semantically related entities, following history-of-ideas research (Lovejoy, 1936; Betti and van den Berg, 2016). HDBSCAN (McInnes et al., 2017) clusters their description embeddings. DEEPSEEK-V4-FLASH names and summarizes every community and cluster from its leading entities. Both levels are built separately for each tradition at four fine-to-coarse resolutions, producing parallel views of the corpus (App. Table 20).

3.3.3 Unified indexing and provenance

A normalized relational schema connects every retrieval level to the primary texts. Each community or cluster links to its member entities and relations. Each entity and relation links to supporting text units, while every text unit records its work, chapter, section, and canonical citation. This creates

a continuous provenance path from a high-level report to citable source passages.

Graph extraction and community or cluster reports use DEEPSEEK-V4-FLASH with thinking enabled and temperature 0.1 (App. F.2). The answer-generation agents use the generator setting in §4.2. One multilingual BGE-M3 encoder (Chen et al., 2024a) ($d=1024$, dense) embeds text units, entities, reports, and queries. For a query q and indexed object x , retrieval uses the inner product $\text{sim}(q, x) = \phi(q)^\top \phi(x)$. The store records each result’s tier t , work match s_w , anchor match s_a , and cosine score s_{cos} . These signals support the ranking in Equation 1, while the tier and source links preserve retrieval provenance. All objects are stored in DuckDB with FAISS inner-product indexes (Johnson et al., 2019).

4 Experiments

4.1 Benchmark construction

To evaluate AI-assisted humanities research, we build a peer-reviewed-paper benchmark in which each paper supplies a research question, a humanist answer, and the primary-source evidence supporting that answer. We collect 406 refereed papers in English or Chinese on classical Chinese and Latin texts from JSTOR, Academia, Google Scholar, and CNKI (Table 3). Two master’s students in history and philology read every paper and manually extract its research question, abstract, findings, and cited primary-source evidence (example in App. E). The records preserve the paper’s wording and finest available locator. Annotators transcribe quoted passages verbatim and record the work, book, chapter, section, or line when the paper supplies a locator. This process yields three evidence granularities: quoted sentences, cited sections, and cited works. The latter two also capture indirect citation through summary, paraphrase, or allusion, motivating the retrieval metrics in Table 4.

Manual annotation captures the flexible structure of humanities articles and their visual evidence cues, including indentation, regular-script Chinese, and italicized Latin. Two annotators independently re-annotated a stratified sample of 100 papers. Question and finding agreement is 0.86, source-work Set F1 is 0.94, normalized-locator agreement is 0.89, and quotation span-overlap F1 is 0.92; a domain expert adjudicated disagreements. At evaluation time, each system receives the paper title and research question and returns an evidence-

Category	Papers	EN/ZH	mean Ev.	mean Pg.
ZH, single	103	56/47	28.8	24
ZH, multi	61	49/12	28.3	26
LA, single	89	55/34	24.3	39
LA, multi	60	51/9	30.8	36
Cross	93	75/18	32.7	33
Total	406	286/120	28.9	27.8

Table 3: Benchmark statistics ($n=406$). ZH/LA = Chinese/Latin tradition; *single/multi* = one/several classics within a tradition, *Cross* = Chinese–Latin comparison; EN/ZH = paper-language; mean Ev./Pg. = mean cited primary-source evidences / page counts. The five categories are the strata for the 100-paper judge study (§4.3).

grounded essay. We assess its direct and indirect evidence use together with its overall scholarly quality.

4.2 Baselines and setup

All systems share the corpus and substrate in Table 2, the BGE-M3 encoder (Chen et al., 2024a) ($d=1024$), the DEEPSEEK-V4-FLASH generator (DeepSeek-AI, 2026) at temperature 0.3, the 406-paper gold records, and the same evaluation code. A robustness run with GEMINI-3-FLASH appears in App. B.1. We compare **SPiRE** with non-agentic and agentic baselines under these matched settings. **Naive LLM** answers from parametric memory. **Text RAG** uses standard text-chunk retrieval followed by generation (Lewis et al., 2020). **GraphRAG** uses its released entity-graph and community-summary retrieval over the same graph (Edge et al., 2025). The two LangChain agents provide generic agentic comparisons. **ReAct** alternates reasoning, tool use, and observation (Yao et al., 2023), while **Plan-Retrieve-Generate (PRG)** decomposes a question into sub-queries, retrieves for each, and synthesizes the results (Ammann et al., 2025). Section 5.3 examines the curation primitives and retrieval tiers, then controls the evidence pool and answer length. **SPiRE** uses its complete seven-agent configuration in the main comparison.

4.3 Metrics and judges

Evidence retrieval. We use IR recall to test whether a system retrieves the same primary-source evidence cited in the reference paper. Following the three citation granularities in Section 4.1, we report sentence recall (sentR), section recall (secR), and work recall (workR). A gold item is recovered when its source work matches and at least one signal fires: a quotation match, a locator match, or BGE-M3 cosine above 0.8 within a language and

0.7 across languages for translated quotations. We also report rank-independent evidence recall (eR) over the full evidence pool available to the final agent. MRR uses the first section-level match, the middle granularity, to measure how early a usable textual locus appears. All recalls are paper-level macro-means. Table 4 reports the 406 papers at $k=10$, and App. Table 10 gives $k \in \{5, 10, 20\}$.

Answer similarity. We report surface similarity between each of the 406 generated essays and the reference paper’s abstract and findings using BGE-M3 cosine, BERTScore (Zhang et al., 2020), and ROUGE-L (Lin, 2004). Full results appear in App. A.

Scholarly judge. Two human experts in classics and philosophy and two LLMs, GPT-5.4 and QWEN3.5-122B-A10B, score each essay from 1 to 5 on AnswerAccuracy, ArgumentDepth, CoverageCompleteness, and EvidenceQuality (Liu et al., 2023; Zheng et al., 2023; Shen et al., 2026). We evaluate a stratified sample of 100 papers, with 20 drawn from each benchmark category in Table 3. Raters receive only the research question and the six anonymised, shuffled system outputs. The rubric therefore accommodates multiple defensible interpretations while applying the same criteria to every system (App. G). We measure agreement with quadratic-weighted Gwet’s AC2 and retain weighted κ as a diagnostic (Gwet, 2008; Feinstein and Cicchetti, 1990).

Cost and equal-volume retrieval. Table 7 reports end-to-end inference cost and a fixed-volume diagnostic that scores only each retriever’s first 50 units. This separates the cost of later curation from retrieval selectivity.

5 Results and Discussion

5.1 Main results

Table 4 reports automatic evidence and answer metrics, while Table 5 presents the scholarly evaluation. Tables 8 and 9 then separate evidence acquisition from evidence curation. A length-controlled evaluation appears in App. Table 11.

5.2 Analysis and discussion

SPiRE leads every baseline in primary-source retrieval across the full 406-paper benchmark (Table 4). Its overall eR is 44.3%, compared with 30.0% for ReAct, 24.4% for PRG, and 22.4% for

System	Evidence retrieval						Answer similarity			
	EvidenceR	EvidenceR@10	WorkR@10	SentenceR@10	SectionR@10	MRR@10	CosSem	ROUGE-L	BERTScore	
Naive LLM	14.3	14.3	17.4	3.6	4.4	15.7	81.0	9.0	85.3	
Text RAG	22.4	14.5	12.5	2.7	3.5	6.5	65.4	3.4	83.8	
GraphRAG	12.5	12.5	13.2	2.3	3.6	6.2	79.0	9.9	84.0	
PRG	24.4	14.3	15.9	3.2	3.1	6.1	71.6	4.5	84.2	
ReAct	30.0	15.4	17.0	3.7	3.5	9.2	63.6	5.6	82.6	
SPIRE (full)	44.3	24.0	42.4	5.6	15.3	33.5	81.8	10.9	84.4	

Table 4: Evidence retrieval and answer similarity at $k=10$ over 406 papers (macro-mean, %). EvidenceR is full-pool recall; MRR@10 uses the first section-level match. Best results are in **bold**.

System	Answer Accuracy				Argument Depth				Coverage Completeness				Evidence Quality			
	GPT	Qwen	H-A	H-B	GPT	Qwen	H-A	H-B	GPT	Qwen	H-A	H-B	GPT	Qwen	H-A	H-B
Naive LLM	3.72	4.01	3.62	4.25	2.81	2.74	2.96	2.06	3.65	3.58	2.51	2.24	2.26	2.46	2.26	1.47
Text RAG	1.78	1.65	1.73	1.76	1.46	1.25	1.17	1.00	1.57	1.45	1.00	1.00	1.55	1.30	1.30	1.00
GraphRAG	2.35	2.55	2.30	2.66	2.37	1.93	2.08	1.40	2.61	2.50	1.85	1.66	1.68	1.39	1.43	1.00
PRG	3.07	2.75	3.41	3.75	2.00	1.52	2.43	2.10	2.26	2.09	2.81	2.73	2.03	1.50	2.45	2.04
ReAct	3.02	2.86	3.20	3.47	2.53	2.20	2.80	2.79	2.94	2.79	3.00	3.20	2.56	2.23	2.77	2.81
SPIRE (full)	4.46	4.95	4.46	4.94	4.63	4.96	3.89	4.57	4.76	4.96	4.39	4.62	4.50	4.90	3.99	4.76

Table 5: Blind scholarly scores on the stratified 100-paper subset (1 to 5). Best results are in **bold**.

Track / statistic	L-L	H-H	H-L	Order
Free track / AC2	.90	.81	.83	SPIRE first
Free track / weighted κ	.89	.59	.69	SPIRE first
Agentic extension / AC2	.92	.89	.69	SPIRE >ReAct>PRG

Table 6: Inter-rater agreement on the judged subset. L/H denote LLM/human raters.

System	min	tokens	calls	cited work@50
Naive LLM	0.22	$\approx 1k$	1	—
Text RAG	1.39	$\approx 36k$	2	49.5
GraphRAG	1.06	$\approx 37k$	2	61.6
PRG	1.12	$\approx 30k$	3	64.3
ReAct	2.72	$\approx 98k$	12	65.8
SPIRE	4.37	$\approx 505k$	22	87.2

Table 7: Per-paper inference cost. Cited work@50 measures whether a cited primary work appears among the first 50 units.

the strongest non-agentic baseline. The advantage holds at every citation granularity. In particular, workR reaches 42.4% versus 17.4%, and sectionR reaches 15.3% versus 4.4%. PRG and ReAct also improve overall evidence recall, confirming the value of agent planning. Naive LLM performs relatively well at work and section levels, likely because these coarser references can be covered by pretrained knowledge. The cases in App. C help explain the lower RAG scores: Text RAG can refuse after off-topic retrieval, while GraphRAG may replace citation-bearing passages with community summaries. **SPIRE** combines local text with community and cluster context, then lets its agents iteratively refine the shared EvidencePool.

Answer similarity shows a different pattern. **SPIRE** achieves the highest cosine similarity and

ROUGE-L, while BERTScore places all six systems within a narrow range. These models can produce topically relevant prose from the research question alone, so embedding similarity may reflect topic overlap more than evidential grounding. GraphRAG exceeds PRG and ReAct on cosine similarity and ROUGE-L despite much lower evidence recall. ReAct shows the reverse pattern, reaching the strongest baseline eR while receiving the lowest cosine similarity and BERTScore. A fluent essay from parametric memory can thus resemble the reference answer as closely as an evidence-grounded essay. This divergence motivates evidence-oriented metrics for AI-assisted humanities research.

The scholarly evaluation reinforces the retrieval results. **SPIRE** leads every aspect for both LLM judges and both human experts, with every score at or above 3.89 (Table 5). Its essays support arguments with original-language quotations, translations, and commentary. The largest margins usually appear in EvidenceQuality and ArgumentDepth, exceeding one point for most raters; App. H gives fragment-level comparisons. GraphRAG remains weak in EvidenceQuality, including a score of 1.00 from Human B. Text RAG also receives several human scores of 1.00, especially for depth and coverage. Naive LLM and the two agentic baselines perform more strongly, with PRG and ReAct gaining most clearly in coverage and evidence quality. On the free track, AC2 is 0.90 for LLM agreement, 0.81 for human agreement, and 0.83 across humans and LLMs. The stronger agreement between model

Configuration	eR	workR	sentR	secR	GPT*	Qwen*
SPIRE (full, 7 primitives)	44.3	42.4	5.6	15.3	3.44	4.01
<i>w/o one agent:</i>						
w/o $\mathcal{N}$	41.9	41.2	4.7	12.7	3.16	3.98
w/o $\mathcal{P}$	42.0	41.8	4.1	13.6	3.18	3.91
w/o $\mathcal{R}$	40.9	40.7	4.3	12.3	3.06	3.93
w/o $\mathcal{S}$	39.1	39.8	3.3	11.1	3.12	3.92
w/o $\mathcal{I}$	27.8	32.4	2.9	10.0	3.07	3.74
<i>Retrieval tier:</i>						
w/o cluster	32.5	31.7	3.6	11.5	3.13	3.77
w/o community	33.5	28.0	3.1	9.9	3.10	3.51
w/o both (local only)	28.5	22.6	2.4	8.1	3.05	3.20

Table 8: Dynamic-pool ablations at $k=10$. GPT/Qwen report mean judge scores; see App. Table 13.

Frozen-pool variant	Used EvidenceR	Claim bindings	GPT*	Qwen*
Full curation	28.6	8.43	4.02	4.15
w/o $\mathcal{N}$	26.8	7.95	3.76	3.89
w/o $\mathcal{P}$	25.9	7.74	3.18	3.36
w/o $\mathcal{R}$	24.1	7.31	3.12	3.22
w/o $\mathcal{S}$	22.7	7.04	3.07	3.17
w/o $\mathcal{I}$	20.5	6.86	2.98	3.08
Direct synthesis	17.9	6.08	2.84	2.94

Table 9: Frozen-pool ablations over all 406 papers. Pool EvidenceR is 39.0; GPT/Qwen are mean judge scores. judges is consistent with their shared systematic preferences (Zheng et al., 2023). Human ratings vary more, reflecting the interpretive range of humanities scholarship. Agreement remains high, and every rater group ranks **SPIRE** first.

This performance carries a computational cost. **SPIRE** uses 4.37 minutes, about 505k tokens, and 22 calls per paper because its agents retrieve, inspect, and revise evidence repeatedly (Table 7). The equal-volume control shows the benefit of this process early in retrieval. Within the first 50 units, **SPIRE** reaches a cited work for 87.2% of papers, compared with rates from 49.5% to 65.8% for the retrieval baselines. The additional computation therefore supports more selective evidence acquisition and more thorough scholarly curation.

The answer-length control gives every system the same 1,500-word target (App. Table 11). **SPIRE** remains strongest on every judge dimension and presents the most reference evidence at a length comparable to the longer baselines.

5.3 Ablation studies

We fix the corpus, substrate, encoder, and generator while ablating the primitives and retrieval tiers (Table 8). Discovering receives the research question, and Representing produces the final essay, so both remain in every configuration. Each other component contributes to evidence recall and judge scores. Local-only retrieval reaches eR=28.5. Adding community or cluster context produces clear gains, while removing the cluster tier from full **SPIRE** lowers eR from 44.3 to 32.5. Intrinsic diagnostics

support this result. Across finer HDBSCAN resolutions, silhouette increases from 0.060 to 0.167 and within-cluster cosine from 0.539 to 0.664, while between-cluster cosine remains near 0.36 and matched random labels remain negative (App. D.1). The clusters supply coherent conceptual neighborhoods beyond explicit graph edges. Removing an agent also changes the shared EvidencePool. The large drop without Illustrating captures both claim-evidence diagnosis and the retrieval triggered by identified gaps.

The frozen-pool control gives every variant the same units, rankings, works, and locators, isolating evidence curation (Table 9). Annotating has the smallest effect. Removing Comparing lowers GPT and Qwen scores by 0.84 and 0.79 despite a moderate change in Used EvidenceR, showing the value of relational organization for argument quality. Removing Referring or Sampling reduces both retained evidence and judge scores because provenance verification and representative selection shape the final essay. Illustrating has the largest post-retrieval effect: its removal lowers Used EvidenceR by 8.1 points and claim-evidence bindings from 8.43 to 6.86. Direct synthesis ranks last throughout. The dynamic and frozen studies reveal complementary functions: iterative agents improve evidence acquisition, while curation agents turn that evidence into a documented argument.

6 Conclusion

We introduce **SPIRE**, a multi-agent framework for evidence-grounded scholarship in the classical humanities. It operationalizes Scholarly Primitives as the unit of computation through seven cooperating agents over a shared EvidencePool and a multi-scale substrate of passages, graph communities, and semantic clusters, making retrieval, contextualization, comparison, and writing inspectable.

On a peer-reviewed-paper benchmark over classical Chinese and Latin scholarship, **SPIRE** achieves higher primary-source evidence retrieval than Naive LLM, Text RAG, GraphRAG, and generic agentic baselines, and receives higher human and LLM judge scores. Its contribution is an architecture making agentic humanities research assistance operational and inspectable, together with a peer-reviewed-paper benchmark and evaluation protocol for further humanities AI research. Future work will explore broader humanities sources and multimodal extensions.

Limitations

Corpora and traditions. We instantiate and evaluate **SPIRE** on classical Chinese and Greco-Roman Latin corpora. These rich and demanding collections support scholarship in history, philosophy, philology, and literature, and provide a strong testbed for evidence-grounded interpretation. Future work can extend the framework to manuscript studies, modern literature, religious studies, archival history, and additional languages and historiographic traditions.

Multimodal evidence. This paper focuses on textual evidence because texts remain foundational to classical studies and many related fields. Humanities research also connects texts with images, archaeological objects, maps, and other material records. Extending **SPIRE** to multimodal retrieval and interpretation is therefore an important next step.

Human-centered scholarship. **SPIRE** supports source discovery, citation checking, comparison, and evidence-grounded synthesis in an inspectable workflow. It can help students learn how to build arguments from sources and help experts explore large collections. Humanistic interpretation remains scholar-led, with final interpretive authority held by researchers.

Ethical Considerations

SPIRE promotes an assistive role for AI in humanities research. Scholarly judgment, teaching responsibility, peer evaluation, and final interpretation remain with researchers. The released datasets are limited to non-commercial research. Copyright remains with the respective authors, publishers, and cultural institutions. Public artifacts follow the original access and licensing conditions, while restricted materials are represented through metadata, derived annotations, or reconstruction scripts. Expert ratings are reported only in aggregate.

Acknowledgments

We thank Zhenqiang Mao for assistance with the historical case analysis. We also thank Dr. Hao Yang and Professor Jingyuan Wu for help organizing the Chinese and Latin historical sources.

AI assistance disclosure. Claude Code with Claude Opus 4.8 assisted with code development and experiment execution. Claude Opus 4.8,

Claude Opus 5.0, and GPT-5.5 were used for language editing. The authors reviewed and verified the resulting code, analyses, and text, and take responsibility for the final paper.

References

- Paul J. L. Ammann, Jonas Golde, and Alan Akbik. 2025. [Question decomposition for retrieval-augmented generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 497–507. Association for Computational Linguistics.
- Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. [Out of one, many: Using language models to simulate human samples](#). *Political Analysis*, 31(3):337–351.
- Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D’Arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, and 6 others. 2024. [Openscholar: Synthesizing scientific literature with retrieval-augmented lms](#). *Preprint*, arXiv:2411.14199.
- Arianna Betti and Hein van den Berg. 2016. [Towards a computational history of ideas](#). In *Proceedings of the Third Conference on Digital Humanities in Luxembourg (CEUR Workshop Proceedings, Vol. 1681)*.
- Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. [Autonomous chemical research with large language models](#). *Nature*, 624:570–578.
- Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. 2024. [Augmenting large language models with chemistry tools](#). *Nature Machine Intelligence*, 6:525–535.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Weize Chen, Ziming You, Ran Li, Yitong Guan, Chen Qian, Chenyang Zhao, Cheng Yang, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2024b. [Internet of agents: Weaving a web of heterogeneous agents for collaborative intelligence](#). *Preprint*, arXiv:2407.07061.
- Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan

- Sun. 2025. [ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery](#). In *International Conference on Learning Representations*.
- DeepSeek-AI. 2026. [DeepSeek-V4: Towards highly efficient million-token context intelligence](#). <https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash>. Preprint, arXiv:2606.19348. Model card. Accessed 26 May 2026.
- Johanna Drucker. 2011. [Humanities approaches to graphical display](#). *Digital Humanities Quarterly*, 5(1).
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. [From local to global: A graph rag approach to query-focused summarization](#). Preprint, arXiv:2404.16130.
- Alvan R. Feinstein and Domenic V. Cicchetti. 1990. [High agreement but low kappa: I. the problems of two paradoxes](#). *Journal of Clinical Epidemiology*, 43(6):543–549.
- Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. [Large language models empowered agent-based modeling and simulation: A survey and perspectives](#). *Humanities and Social Sciences Communications*, 11(1):1259.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. [Retrieval-augmented generation for large language models: A survey](#). Preprint, arXiv:2312.10997.
- Alireza Ghafarollahi and Markus J. Buehler. 2024. [Sci-Agents: Automating scientific discovery through multi-agent intelligent graph reasoning](#). Preprint, arXiv:2409.05556.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. [DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning](#). *Nature*, 645:633–638.
- Kilem Li Gwet. 2008. [Computing inter-rater reliability and its variance in the presence of high agreement](#). *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.
- Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A. Rossi, Subhabrata Mukherjee, Xianfeng Tang, Qi He, Zhigang Hua, Bo Long, Tong Zhao, Neil Shah, Amin Javari, Yinglong Xia, and Jiliang Tang. 2025. [Retrieval-augmented generation with graphs \(GraphRAG\)](#). Preprint, arXiv:2501.00309.
- Yichen He, Guanhua Huang, Peiyuan Feng, Yuan Lin, Yuchen Zhang, Hang Li, and Weinan E. 2025. [PaSa: An LLM agent for comprehensive academic paper search](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11663–11679. Association for Computational Linguistics.
- John J. Horton, Apostolos Filippas, and Benjamin S. Manning. 2023. [Large language models as simulated economic agents: What can we learn from homo silicus?](#) Working Paper 31122, National Bureau of Economic Research.
- Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. [SWE-bench: Can language models resolve real-world GitHub issues?](#) In *International Conference on Learning Representations*.
- Matthew L. Jockers. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. [Billion-scale similarity search with GPUs](#). *IEEE Transactions on Big Data*, 7(3):535–547.
- Paul Kelly, Jonathan Schild, and Amir Jafari. 2025. [FolkRAG: A retrieval-augmented generation system for cultural heritage materials](#). *Neural Computing and Applications*.
- DaeHo Kim, SoYeop Yoo, and OkRan Jeong. 2026. [MedSumGraph: Enhancing GraphRAG for medical QA with summarization and optimized prompts](#). *Artificial Intelligence in Medicine*, 172:103311.
- Simon Knollmeyer, Oğuz Caymaz, and Daniel Großmann. 2025. [Document GraphRAG: Knowledge graph enhanced retrieval augmented generation for document question answering within the manufacturing domain](#). *Electronics*, 14(11):2102.
- Julia Kristeva. 1980. *Desire in Language: A Semiotic Approach to Literature and Art*. Columbia University Press.
- Jakub Lála, Odhran O’Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodrigues, and Andrew D. White. 2023. [Paperqa: Retrieval-augmented generative agent for scientific research](#). Preprint, arXiv:2312.07559.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.
- Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2023. [Econagent: Large language model-empowered agents for simulating macroeconomic activities](#). *arXiv preprint arXiv:2310.10436*.

- Ruochen Li, Teerth Patel, Qingyun Wang, and Xinya Du. 2024. [Mlr-copilot: Autonomous machine learning research based on large language models agents](#). *Preprint*, arXiv:2408.14033.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-Eval: NLG evaluation using GPT-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics.
- Arthur O. Lovejoy. 1936. *The Great Chain of Being: A Study of the History of an Idea*. Harvard University Press, Cambridge, Massachusetts.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. [The ai scientist: Towards fully automated open-ended scientific discovery](#). *Preprint*, arXiv:2408.06292.
- Haoran Luo, Haihong E, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Zemin Kuang, Meina Song, Yifan Zhu, and Luu Anh Tuan. 2025. [HyperGraphRAG: Retrieval-augmented generation via hypergraph-structured knowledge representation](#). In *Advances in Neural Information Processing Systems*.
- Leland McInnes, John Healy, and Steve Astels. 2017. [hdbscan: Hierarchical density based clustering](#). *Journal of Open Source Software*, 2(11):205.
- Franco Moretti. 2013. *Distant Reading*. Verso.
- Keerthana Murugaraj, Salima Lamsiyah, Marten Durig, and Martin Theobald. 2025. [Topic-RAG for historical newspapers: Enhancing information retrieval in humanities research through topic-based retrieval-augmented generation](#). *Computational Humanities Research*, 1.
- Carole L. Palmer, Lauren C. Teffeau, and Carrie M. Pirman. 2009. Scholarly information practices in the online environment: Themes from the literature and implications for library service development. Technical report, OCLC Research.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simula-
cra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.
- Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*.
- Geoffrey Rockwell and Stéfan Sinclair. 2016. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. MIT Press.
- Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. 2024. [Mathematical discoveries from program search with large language models](#). *Nature*, 625:468–475.
- Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. 2025. [Agent laboratory: Using LLM agents as research assistants](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5977–6043. Association for Computational Linguistics.
- William F. Shen, Xinchu Qiu, Chenxi Whitehouse, Lisa Alazraki, Shashwat Goel, Francesco Barbieri, Timon Willi, Akhil Mathur, and Ilias Leontiadis. 2026. [Re-thinking rubric generation for improving LLM judge and reward modeling for open-ended tasks](#). *Preprint*, arXiv:2602.05125.
- Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*.
- Quentin Skinner. 1969. [Meaning and understanding in the history of ideas](#). *History and Theory*, 8(1):3–53.
- Barbara Herrnstein Smith. 2016. [What was “close reading”? a century of method in literary studies](#). *The Minnesota Review*, 2016(87):57–75.
- The Latin Library. 2026. The Latin Library. <https://www.thelatinlibrary.com/>. Accessed 26 May 2026.
- Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. 2019. [From Louvain to Leiden: Guaranteeing well-connected communities](#). *Scientific Reports*, 9(1):5233.
- Ted Underwood. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.
- John Unsworth. 2000. Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this? In *Humanities Computing: Formal Methods, Experimental Practice*. King’s College London.
- Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. 2025. [Medical graph RAG: Evidence-based](#)

medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28443–28467, Vienna, Austria. Association for Computational Linguistics.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xiong-Hui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Li, Yuyu Luo, and Chenglin Wu. 2025. AFlow: Automating agentic workflow generation. In *International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*.

Tianshi Zheng, Zheyang Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From automation to autonomy: A survey on large language models in scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17733–17750, Suzhou, China. Association for Computational Linguistics.

Gucheng Zhou, editor. 1992. *Zhongguo Xueshu Mingzhu Tiyao (Compendium of Chinese Academic Masterworks)*. Fudan University Press, Shanghai, China. Multi-volume series published from 1992 to 1999.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. 2024. GPTSwarm: Language agents as optimizable graphs. In *International Conference on Machine Learning*.

A Full Results

This section expands the main results along three dimensions. It reports the complete retrieval sweep,

controls essay length under one shared instruction, and separates performance across the five benchmark categories. Together, these analyses show whether the main comparison holds across retrieval depth, output length, and research setting.

A.1 Retrieval and answer-similarity sweep ($k \in \{5, 10, 20\}$)

Table 10 expands Table 4 to $k \in \{5, 10, 20\}$. **SPIRE** leads every retrieval metric at each depth, and its advantage remains clear as more results enter the evidence set. The three answer similarity metrics are reported once at $k=10$. Their ordering differs from both evidence recovery and scholarly judging, which supports their use as descriptive measures rather than grounding measures (§4.3).

A.2 Length-controlled evaluation

We append the same instruction to every system prompt, asking for a final essay of approximately 1,500 words. We then compute evidence recall over the presented evidence and apply the four scholarly criteria on the same stratified 100-paper subset. Visible output length is counted in Unicode characters so that Chinese and English responses share one measure. Table 11 reports the results.

Under the shared target, **SPIRE** remains within 154 characters of ReAct, 195 of GraphRAG, and 451 of Naive LLM, with PRG in the same range. It leads every judge dimension by 0.79 to 1.33 points over the strongest corresponding baseline. It also presents the most reference evidence. The result confirms that its advantage persists at comparable answer length.

A.3 Per-category breakdown (retrieval and answer similarity)

Table 12 separates the $k=10$ results across the five categories in Table 3. **SPIRE** leads EvidenceR in every category, ranging from 36.5 to 52.6. The generic agents extend evidence coverage beyond the non-agentic baselines. ReAct has the strongest baseline EvidenceR in all five categories, while PRG usually falls between ReAct and the non-agentic systems.

The remaining gap is concentrated at finer source levels. On cross-tradition questions, PRG and ReAct reach 17.7 and 21.4 EvidenceR, compared with 36.5 for **SPIRE**. Their WorkR scores are 8.4 and 10.6, while **SPIRE** reaches 33.9. Neither generic agent exceeds 6.8 SectionR in any category; **SPIRE** reaches 19.6 and 21.2 on the two

System	k	EvidenceR	EvidenceR@ k	WorkR	SentenceR	SectionR	MRR	CosSem	ROUGE-L	BERTScore
Naive LLM	5	–	13.9	16.8	3.4	4.2	15.6	–	–	–
	10	14.3	14.3	17.4	3.6	4.4	15.7	81.0	9.0	85.3
	20	–	14.3	17.4	3.6	4.4	15.7	–	–	–
Text RAG	5	–	10.6	9.2	2.1	2.5	5.8	–	–	–
	10	22.4	14.5	12.5	2.7	3.5	6.5	65.4	3.4	83.8
	20	–	17.5	16.1	3.5	4.5	6.8	–	–	–
GraphRAG	5	–	11.8	12.3	2.2	3.2	5.8	–	–	–
	10	12.5	12.5	13.2	2.3	3.6	6.2	79.0	9.9	84.0
	20	–	12.5	13.2	2.3	3.6	6.2	–	–	–
PRG	5	–	9.9	11.1	2.3	1.6	5.3	–	–	–
	10	24.4	14.3	15.9	3.2	3.1	6.1	71.6	4.5	84.2
	20	–	18.3	22.6	4.5	5.7	6.7	–	–	–
ReAct	5	–	11.6	11.9	2.5	2.4	8.4	–	–	–
	10	30.0	15.4	17.0	3.7	3.5	9.2	63.6	5.6	82.6
	20	–	19.4	23.1	5.1	5.6	9.7	–	–	–
SPIRE (full)	5	–	19.3	38.5	4.1	10.1	32.7	–	–	–
	10	44.3	24.0	42.4	5.6	15.3	33.5	81.8	10.9	84.4
	20	–	28.3	45.1	7.8	18.6	33.8	–	–	–

Table 10: Retrieval sweep on all 406 papers. EvidenceR and answer similarity appear at $k=10$ because they are independent of retrieval depth. MRR uses the first section-level match. Bold marks the best retrieval result at $k=10$.

System	Visible chars	Presented evidence			GPT-5.4				Qwen3.5			
		EvidenceR@10	WorkR@10	SectionR@10	Acc	Dep	Cov	EQ	Acc	Dep	Cov	EQ
Naive LLM	5,681	22.0	21.2	7.4	3.55	3.05	3.75	2.81	3.80	2.87	3.75	2.74
Text RAG	2,615	19.8	12.5	3.3	2.09	1.80	1.73	2.21	1.85	1.45	1.64	1.65
GraphRAG	5,937	16.4	12.7	6.1	2.54	2.30	3.09	1.68	1.89	1.53	1.95	1.00
PRG	5,437	19.9	13.9	3.0	3.86	3.78	3.92	3.76	3.67	3.20	3.58	3.17
ReAct	6,286	23.7	22.0	5.2	3.63	3.69	3.63	3.76	3.55	3.27	3.46	3.22
SPIRE	6,132	34.6	51.1	23.4	4.65	4.76	4.77	4.66	4.63	4.60	4.71	4.45

Table 11: Length-controlled results on the judged 100-paper subset. Every system receives the same 1,500-word target.

Chinese categories. Planning and repeated retrieval therefore broaden coverage, while the scholarly workflow more consistently localizes evidence for citation. ReAct’s FindingsCos also falls to 0.564 and 0.488 on the two Latin categories. GraphRAG remains the one exception at SectionR for multi-work Latin questions, scoring 8.1 versus 7.3 for **SPIRE**.

A.4 Component ablations

Table 13 is an end-to-end, dynamic-pool ablation: removing an agent may alter both the shared EvidencePool and later reflective retrieval. The separate frozen-pool control in Table 9 runs Discovering once, freezes every unit, ranking, work, and locator, disables secondary retrieval, and then changes only post-retrieval curation. The dynamic design measures upstream pool changes; the frozen design isolates post-retrieval curation.

B Robustness and Prompting Controls

B.1 Generator robustness

We repeat the comparison with GEMINI-3-FLASH as the generator while holding the corpus, substrate, BGE-M3 encoder, retrieval protocol, and rubric judges fixed (§4.2). Tables 14 and 15 report the Gemini results. The corresponding DEEPSEEK-V4-FLASH results appear in Tables 4 and 5. **SPIRE** retains the leading position in retrieval and on every judge dimension. Answer similarity remains closely grouped.

Retrieval. Under GEMINI-3-FLASH, **SPIRE** keeps a wide retrieval margin over every baseline. Its full-pool EvidenceR is 54.1, compared with 31.6 for ReAct and 25.1 for PRG, while WorkR is 51.2 versus 18.4 and 16.6. The seven-primitive workflow retains its advantage in provenance-aware acquisition under both generators.

Answer similarity. Among the four Gemini systems with available scores, **SPIRE**, Naive LLM, and GraphRAG fall within 1.4 points on findings

Category	System	EvidenceR	WorkR	SectionR	MRR	SourceSemR	FindingsCos
ZH, single (LOCAL_ZH)	SPIRE	52.6	43.1	19.6	34.0	66.2	0.829
	PRG	33.1	15.5	0.9	2.8	65.1	0.742
	ReAct	37.6	19.5	2.5	8.8	65.6	0.768
	Text RAG	30.1	8.7	0.3	1.0	64.3	0.685
	GraphRAG	16.0	11.9	1.3	2.3	61.9	0.794
	Naive LLM	21.9	17.7	2.7	9.6	64.0	0.823
ZH, multi (GLOBAL_ZH)	SPIRE	43.7	35.6	21.2	35.9	63.4	0.816
	PRG	22.2	10.6	1.3	2.6	62.6	0.709
	ReAct	26.1	10.0	1.2	3.2	62.8	0.726
	Text RAG	18.9	4.2	0.5	0.9	62.3	0.655
	GraphRAG	8.9	2.1	0.1	2.3	59.7	0.773
	Naive LLM	13.0	9.2	4.7	10.3	61.3	0.804
LA, single (LOCAL_LA)	SPIRE	45.5	55.5	12.2	35.0	60.4	0.823
	PRG	25.6	27.1	6.5	10.6	59.8	0.720
	ReAct	32.8	24.5	6.8	14.4	60.2	0.564
	Text RAG	23.7	24.8	7.1	12.7	60.0	0.658
	GraphRAG	15.6	25.6	6.4	11.2	58.4	0.806
	Naive LLM	14.3	31.4	7.2	26.5	58.8	0.810
LA, multi (GLOBAL_LA)	SPIRE	40.4	41.7	7.3	31.9	59.5	0.805
	PRG	20.6	17.2	2.6	7.7	59.7	0.716
	ReAct	29.9	18.7	5.2	14.2	59.8	0.488
	Text RAG	24.6	17.5	5.6	14.6	59.4	0.633
	GraphRAG	15.8	18.0	8.1	10.7	58.7	0.792
	Naive LLM	15.7	20.9	6.7	26.0	58.0	0.801
Cross (ZH-LA)	SPIRE	36.5	33.9	16.0	31.0	60.5	0.811
	PRG	17.7	8.4	3.4	6.1	60.4	0.689
	ReAct	21.4	10.6	1.3	4.0	60.0	0.595
	Text RAG	13.8	7.4	3.6	3.4	59.8	0.628
	GraphRAG	5.7	6.8	2.1	4.2	57.4	0.778
	Naive LLM	5.8	6.8	1.4	6.6	58.5	0.805

Table 12: Results by benchmark category at $k=10$. SourceSemR measures semantic recall over gold source evidence, and FindingsCos compares each essay with the paper findings. Bold marks the best retrieval result by category.

Configuration	k	EvidenceR	EvidenceR@ k	WorkR	SentenceR	SectionR	MRR	CosSem	ROUGE-L	BERTScore	GPT*	Qwen*
SPIRE (full, 7 primitives)	5	—	19.3	38.5	4.1	10.1	32.7	—	—	—	—	—
	10	44.3	24.0	42.4	5.6	15.3	33.5	81.8	10.9	84.4	3.44	4.01
	20	—	28.3	45.1	7.8	18.6	33.8	—	—	—	—	—
<i>w/o one agent (removed from the workflow):</i>												
w/o Annotating ($\mathcal{N}$)	5	—	19.6	38.1	3.0	9.6	31.1	—	—	—	—	—
	10	41.9	23.3	41.2	4.7	12.7	31.8	82.2	10.9	84.4	3.16	3.98
	20	—	27.5	44.4	6.4	16.0	32.1	—	—	—	—	—
w/o Comparing ($\mathcal{P}$)	5	—	18.8	38.3	2.4	8.7	32.7	—	—	—	—	—
	10	42.0	23.1	41.8	4.1	13.6	33.6	81.9	10.9	84.4	3.18	3.91
	20	—	27.4	43.3	4.9	17.0	33.8	—	—	—	—	—
w/o Referring ($\mathcal{R}$)	5	—	18.6	37.4	2.6	8.8	35.6	—	—	—	—	—
	10	40.9	22.4	40.7	4.3	12.3	36.5	82.0	11.0	84.5	3.06	3.93
	20	—	26.5	43.3	6.0	15.5	36.8	—	—	—	—	—
w/o Sampling ($\mathcal{S}$)	5	—	16.7	36.7	2.5	8.5	34.0	—	—	—	—	—
	10	39.1	20.2	39.8	3.3	11.1	34.9	82.0	10.8	84.4	3.12	3.92
	20	—	23.9	42.8	4.8	17.1	35.2	—	—	—	—	—
w/o Illustrating ($\mathcal{I}$)	5	—	12.1	29.8	1.9	7.1	21.4	—	—	—	—	—
	10	27.8	14.6	32.4	2.9	10.0	21.9	82.0	10.9	84.4	3.07	3.74
	20	—	17.3	34.4	3.9	12.6	22.1	—	—	—	—	—
<i>Retrieval tier disabled:</i>												
w/o cross-context cluster	5	—	14.0	28.3	1.8	7.9	26.4	—	—	—	—	—
	10	32.5	16.9	31.7	3.6	11.5	27.1	70.9	8.1	83.0	3.13	3.77
	20	—	20.8	34.4	4.7	13.7	27.3	—	—	—	—	—
w/o intra-context community	5	—	10.7	25.7	1.7	7.0	23.5	—	—	—	—	—
	10	33.5	12.8	28.0	3.1	9.9	24.1	63.4	5.7	82.2	3.10	3.51
	20	—	16.4	31.1	4.2	12.2	24.3	—	—	—	—	—
w/o both (local only)	5	—	7.5	20.6	1.7	6.0	22.0	—	—	—	—	—
	10	28.5	9.4	22.6	2.4	8.1	22.5	63.3	5.9	82.2	3.05	3.20
	20	—	11.3	25.5	3.0	9.6	22.7	—	—	—	—	—

Table 13: Dynamic-pool ablations across three retrieval depths. EvidenceR, answer similarity, and mean judge scores appear at $k=10$ because they are independent of retrieval depth.

cosine and BERTScore. Surface similarity again compresses systems whose evidence retrieval differs substantially.

LLM judge. **SPIRE** ranks first on every rubric axis under both Gemini judge runs (Table 15). Its overall scores are 4.21/4.58, versus 3.83/3.88 for PRG and 3.67/3.85 for ReAct. Its Evidence-Quality scores are 4.20/4.55, compared with the strongest baseline scores of 3.80/3.82.

B.2 Chain-of-thought prompting control

We add explicit chain-of-thought (CoT) instructions to Naive LLM and Text RAG, holding model, temperature, retrieval, generation budget, and evaluation fixed. CoT modestly improves judged synthesis but affects retrieval inconsistently: it lowers Naive eR and raises Text RAG eR by 1.9 points (Table 16). Both variants remain below **SPIRE** on evidence recovery and judged quality.

C Failure Analysis

The two expert judges reviewed all 2,236 gold evidence instances in the stratified 100-paper set and coded every unrecovered instance. They also annotated answer and expression problems in representative cases. The resulting taxonomy covers textual localization, interpretation, and scholarly expression (Table 17). The Share column divides only the unrecovered gold evidence. It does not measure the proportion of essays with errors. The listed evidence categories sum to 99.1%, with the remaining 0.9% assigned to title variants of the same work. Answer and expression rows identify case types and therefore carry no percentage.

Most evidence misses are localization rather than total grounding failure. In GLOBAL_ZH_007, **SPIRE** retrieves 19 passages from the *Xunzi* but stays in chapters 1–9 while the paper cites chapters 19 and 23. Alternative evidence captures valid plural paths: in GLOBAL_LA_061 the paper uses both Ennius and Livy, while **SPIRE** misses Ennius but uses the paper’s supportive Livy passage. True misses often arise from underspecified scope. For GLOBAL_ZH_078, the broad question never names a source, so provenance-first retrieval fails to anchor the *Analects* evidence selected by the author.

SPIRE, case 1: period/term conflation. On the cross-tradition question H004 (“parallels between Epictetus and *classical* Chinese philosophy”), **SPIRE** grounds the Chinese side in Song–Yuan Neo-Confucian sources rather than the pre-Qin classics the field denotes by “classical”:

In the *Song-Yuan xue'an* (Records of Song and Yuan Scholars), a passage attributed to the Song Confucian Wang Kaizu draws a ... distinction between different modes of attending to the heart-mind (*xin*)... [E1]

Expert A flagged exactly this: many cited sources came from the Song–Ming period, so the judgement about “classical Chinese philosophy” may be problematic. The cause is the generator’s parametric knowledge, which blends historically distinct Confucian strata (pre-Qin Confucius/Mencius vs. Song–Ming *lixue*); a period-aware retrieval constraint or a domain authority list pinning “classical” to the intended sub-corpus would mitigate it.

SPIRE, case 2: self-translated long Latin quotations. For long Latin passages **SPIRE** supplies its *own* English rendering rather than an established scholarly translation. On H004 it quotes Aulus Gellius’s *Noctes Atticae* at length:

“Natura” inquit “omnium rerum, quae nos genuit, induit nobis inoleuitque ... amorem nostri et caritatem...” (*Noctes Atticae*) [E5]

and renders it in its own prose (“nature instills in every creature ... an innate love and care for itself”). Humanities citation norms expect the canonical published translation (e.g. the Loeb Classical Library edition) for such passages because they are accurate, vetted, and citable. Pairing the retrieved primary text with a registered translation resource would resolve this.

GraphRAG: secondary literature substituted for the primary text (severe). On H008 (Daoism, Confucianism, and Augustine on learning), asked about the *Daoist* position, GraphRAG never reaches Laozi or Zhuangzi; it instead retrieves a Neo-Confucian critic and a Qing philologist, and imports Buddhism the question never raised:

The retrieved passage from Zhang Zai criticizes Buddhism and Daoism. ... The passage from Dai Zhen notes that Laozi said “abandon learning and be free of worry”...

Expert B scored EvidenceQuality = 1 and noted that the Daoist side used neither Zhuangzi nor direct Daoist evidence, but secondary literature from Zhang Zai and even introduced Buddhism. Community-summary retrieval surfaces commentary *about* the topic rather than the cited primary source.

System	Evidence retrieval						Answer similarity		
	EvidenceR	EvidenceR@10	WorkR@10	SentenceR@10	SectionR@10	MRR@10	CosSem	ROUGE-L	BERTScore
Naive LLM	14.0	14.0	16.8	3.5	4.3	15.2	80.0	7.7	85.1
Text RAG	19.5	14.2	10.4	2.2	3.1	5.4	63.6	3.2	82.7
GraphRAG	16.4	16.4	12.8	2.5	6.1	7.6	79.4	8.8	83.9
PRG	25.1	15.8	16.6	3.3	3.6	7.4	—	—	—
ReAct	31.6	17.2	18.4	4.0	4.2	10.5	—	—	—
SPIRE (full)	54.1	34.9	51.2	8.4	23.1	52.7	81.4	10.2	84.3

Table 14: Retrieval and answer similarity with GEMINI-3-FLASH. Bold marks the best retrieval result. Answer similarity covers Naive LLM, Text RAG, GraphRAG, and **SPIRE**.

System	Rater A (GPT-5.4)					Rater B (Qwen3.5)				
	Acc	Dep	Cov	EQ	Ov	Acc	Dep	Cov	EQ	Ov
SPIRE	4.25	4.03	4.34	4.20	4.21	4.59	4.56	4.60	4.55	4.58
PRG	3.96	3.55	3.99	3.80	3.83	4.13	3.66	3.97	3.77	3.88
ReAct	3.73	3.46	3.70	3.80	3.67	4.01	3.73	3.85	3.82	3.85
Naive LLM	3.90	3.13	3.96	2.31	3.33	3.35	3.32	3.15	2.50	3.08
Text RAG	3.17	3.09	3.12	2.68	3.02	3.37	3.15	3.25	3.17	3.24
GraphRAG	2.60	2.24	2.86	1.42	2.28	2.67	2.04	2.64	2.24	2.40

Table 15: Free-track LLM judge scores with GEMINI-3-FLASH on the same 100 questions. Ov is the mean of the four aspects. Bold marks the best result.

System	eR	eR@10	workR	sentR	secR	GPT overall	Qwen overall
Naive LLM	14.3	14.3	17.4	3.6	4.4	3.11	3.20
Naive LLM + CoT	12.6	12.6	15.1	3.2	3.8	3.45	3.48
Text RAG	22.4	14.5	12.5	2.7	3.5	1.59	1.41
Text RAG + CoT	24.3	15.2	14.3	3.1	3.6	1.75	1.55
SPIRE	44.3	24.0	42.4	5.6	15.3	4.59	4.94

Table 16: CoT prompting control on all 406 papers. Retrieval metrics are paper-level macro-means (%); judge columns average the four rubric dimensions on the 100-paper judged subset.

Level	Category	Share	Example
Evidence	Granularity gap (right work, wrong locus)	57.7%	GLOBAL_ZH_007
Evidence	True miss (no primary work found)	18.8%	GLOBAL_ZH_078
Evidence	Alternative supportive evidence	15.0%	GLOBAL_LA_061
Evidence	Adjacent-context miss	4.0%	GLOBAL_LA_054
Evidence	Source-layer shift	3.6%	LOCAL_LA_087
Answer	Scope drift	case	CROSS_018
Answer	Argument gap	case	GLOBAL_ZH_011
Answer	Coverage imbalance	case	CROSS_009
Expression	Non-canonical translation	case	GLOBAL_LA_060
Expression	Truncated quotation	case	GLOBAL_LA_048

Table 17: Error taxonomy on the 100-paper expert subset. Evidence shares are percentages of unrecovered gold evidence, rather than percentages of essays.

Text RAG: off-topic retrieval, then refusal (severe). On CROSS_006 (“should Confucius be regarded as a *Sinicus Cicero*?”), flat cosine retrieval returns only unrelated Latin rhetoric (Quintilian’s *Institutio Oratoria*, Seneca’s *De Brevitate Vitae* and *Epistulae*), with no material on Confucius or the *Confucius Sinarum Philosophus*, so the answer collapses to a single sentence:

“The retrieved passages do not contain any information about Confucius . . . Therefore, they are insufficient to answer whether Confucius should be regarded as a *Sinicus Cicero*.”

When retrieval misses, the flat baseline neither supplements nor reasons; its answers collapse to such refusals (median length ~ 270 characters), which the judge penalises on every axis.

Naive LLM: ungrounded recall (severe). With no retrieval, the no-retrieval baseline answers CROSS_001 (Confucius vs. Cicero) fluently from parametric memory and cites from recollection:

{source: *Lunyu / Analects* 12.1, “to master oneself and return to ritual is *ren*”; role: central; confidence: high}; {source: *De officiis*. . . }

The named loci are plausible but unverifiable against the corpus and prone to misattribution. This failure motivates evidence-grounded retrieval.

C.1 Failure modes of generic agent baselines

PRG and ReAct use the same retrieval substrate and generator as **SPIRE**. Their paper-level diagnostics therefore help locate the effect of the agent workflow. Table 18 summarizes two distinct patterns.

System	Diagnostic	Mean units	Overall	Cross
PRG	Pool judged irrelevant	50.0	34.2%	51.6%
ReAct	No retrieved unit cited	135.9	15.4%	21.6%
ReAct	Gold work in pool, none named	135.9	4.9%	n/a

Table 18: Paper-level evidence-use diagnostics for the generic agents. Mean units is the retrieved volume per paper.

PRG: useful evidence judged irrelevant. PRG retrieves a fixed pool of 50 units before synthesis. On CROSS_006, which asks whether Confucius should be regarded as a *Sinicus Cicero*, that pool covers all gold evidence in the reference paper, with EvidenceR = 100.0. Its answer nevertheless concludes that the retrieved texts provide no relevant evidence for the comparison. The missing step is alignment: the Latin and Chinese passages must be treated as two sides of the same question. This pattern appears in 51.6% of cross-tradition papers, compared with 24.3% of single-work Chinese papers. It is the type of evidence decision supported by the Comparing primitive.

ReAct: retrieval without consistent binding. ReAct retrieves 135.9 units per paper on average. The larger pool expands coverage, yet 15.4% of its essays cite none of the retrieved units. In another 4.9%, the pool contains a gold work that the essay never names. The zero-citation rate rises to 21.6% for cross-tradition questions, where evidence from both sides must be aligned before synthesis. This pattern corresponds to the role of Illustrating, which links claims to supporting passages.

The category results show the same progression at finer granularity. On cross-tradition questions, PRG and ReAct reach EvidenceR scores of 17.7 and 21.4, but their WorkR scores are 8.4 and 10.6, and their SectionR scores are 3.4 and 1.3 (Table 12). Generic planning broadens the pool. Scholarly primitives more consistently turn that pool into localized, citable evidence.

The taxonomy therefore qualifies the aggregate eR result. **SPIRE** still has true retrieval failures and scope errors, but the largest category is a locator-level gap and some strict misses remain defensible alternative paths. The baseline cases show a different concentration of failure: wrong source layers, unusable retrieval, and unsupported from-memory citation. These patterns explain why work/section recall and EvidenceQuality separate the systems more clearly than surface similarity.

D Graph, Community and Cluster Substrate

D.1 Intrinsic semantic-cluster evaluation

We evaluate HDBSCAN in the original 1,024-dimensional embedding space against cluster-count-matched KMeans and random-label controls. HDBSCAN has the highest silhouette at every evaluated scale (Table 19). Figure 3 visualizes the resulting cluster geometry. As minimum cluster size decreases, within-cluster cosine rises from 0.539 to 0.664 while between-cluster cosine stays near 0.36. This intrinsic separation complements the downstream result: removing the cluster tier lowers eR from 44.3 to 32.5.

D.2 Graph and cluster artefacts per scale

Table 20 confirms the design rationale of §3.3: even on the combined cross corpus the mean cross-language balance of the coarsest, most reliable clusters is 0.008. The vector index captures cross-context recurrence; the Comparing primitive performs cross-tradition alignment at the agent level.

E Benchmark Record and Extracted-Graph Examples

Benchmark record. A representative gold record (one of 406; §4.1). Internal identifiers and annotation labels are omitted; we keep the fields a system is scored against.

Gold benchmark record (abridged)

```

1 paper: "Pietas in pro Sexto Roscio of Cicero and
  Confucian xiao"
2 paper_language: en literature_languages:
  zh_classical + la_classical
3
4 research_question:
5   How does Cicero's pietas in pro Sexto Roscio
  relate to Confucian xiao,
6   and what does comparison show about filial duty
  and social order?
7
8 research_findings (excerpt):
9   Parricide is treated as the ultimate filial
  impiety; Cicero rejects the
10  charge by arguing ex persona, ex causa, and ex
  facto ipso.
11
12 evidences (2 of 13 shown):
13   E01 Pro Roscio Amerino loc 37
  translated_quote central
14   "Sextus Roscius stands accused of the
  murder of his father. ..."
15   E04 Pro Roscio Amerino loc 69 original_quote
  supporting
16   "quod in impios singulare supplicium
  invenerunt"
```

Extracted graph. Selected fields from per-chunk DEEPSEEK-V4-FLASH output (§3.3) are shown below. The arrays are shortened after schema vali-

Min. size	Coverage	HDBSCAN sil.	KMeans sil.	Random sil.	Intra cos	Inter cos
100	24.4%	0.060	0.053	−0.060	0.539	0.352
30	26.1%	0.077	0.036	−0.140	0.589	0.355
10	30.4%	0.093	−0.051	−0.241	0.643	0.360
3	41.4%	0.167	n/a	−0.262	0.664	0.370

Table 19: Intrinsic evaluation of semantic clusters. Coverage is the percentage of embedded entities assigned to non-noise clusters.

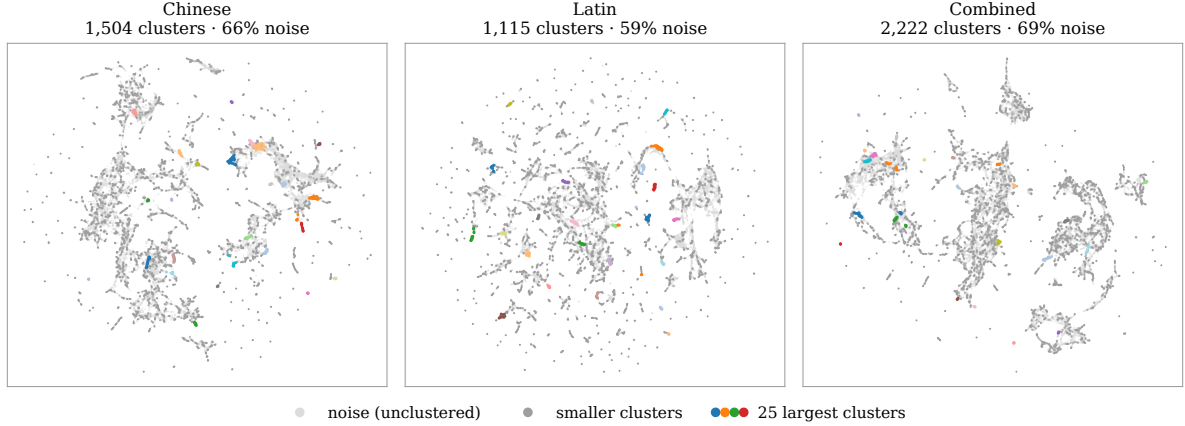


Figure 3: UMAP projections of the HDBSCAN semantic-cluster spaces for the Chinese, Latin, and combined cross-context indices at minimum cluster size 10 (left to right). Panel titles report cluster counts and noise shares. Colored points mark the 25 largest clusters; dark gray points are smaller non-noise clusters, and light gray points are HDBSCAN noise. UMAP is used only for visualization; Table 19 evaluates clusters in the original 1,024-dimensional embedding space.

Lang	Algo	Scale	#groups	max	avg sz	x-lang bal
zh	Leiden	res=2.0	1,330	6,218	81.0	−
zh	Leiden	res=8.0	1,562	3,451	69.0	−
zh	Leiden	res=20.0	2,072	2,356	52.0	−
zh	Leiden	res=30.0	2,488	2,290	43.3	−
zh	HDBSCAN	mcs=100	116	1,684	315.9	0.000
zh	HDBSCAN	mcs=30	422	629	91.9	0.000
zh	HDBSCAN	mcs=10	1,504	395	29.4	0.000
zh	HDBSCAN	mcs=3	7,753	59	7.4	0.000
la	Leiden	res=2.0	776	5,129	80.4	−
la	Leiden	res=8.0	949	1,456	65.7	−
la	Leiden	res=20.0	1,319	1,199	47.3	−
la	Leiden	res=30.0	1,603	1,139	38.9	−
la	HDBSCAN	mcs=100	68	1,663	337.5	0.000
la	HDBSCAN	mcs=30	291	607	89.9	0.000
la	HDBSCAN	mcs=10	1,115	211	27.7	0.000
la	HDBSCAN	mcs=3	5,006	45	7.6	0.000
cross	HDBSCAN	mcs=100	147	2,472	346.5	0.008
cross	HDBSCAN	mcs=30	563	854	96.8	0.020
cross	HDBSCAN	mcs=10	2,222	367	28.6	0.059
cross	HDBSCAN	mcs=3	11,958	66	7.2	0.169

Table 20: Offline graph/cluster artifacts per scale. “x-lang bal” is the mean cross-language balance of HDBSCAN clusters; it is near zero even on the combined cross corpus, empirically justifying agent-level cross-tradition alignment (§3.3).

dition; Chinese glyphs are romanized to match the paper’s convention.

Extracted graph – Latin chunk (Ammianus Marcellinus, Res Gestae)	
1	{ "entities": {
2	{ "type": "Person", "title_modern": "Ammianus Marcellinus",
3	"description_modern": "Roman historian and author of the Res Gestae.",
4	"evidence_spans": [1, 9],

```

5  { "type": "Person", "title_modern": "Nohodares",
6    "description_modern": "A Persian nobleman who
7    planned an invasion of Mesopotamia.",
8    "evidence_spans": [2] },
9  { "type": "Concept", "domain": "Politics &
10  Governance", "title_modern": "Saracen customs",
11    "description_original": "Saracenorum irruptiones
12    et mores",
13    "description_modern": "The nomadic, predatory
14    life and military habits of the Saracens.",
15    "evidence_spans": [9, 10, 11, 12, 13, 14, 15] },
16  { "type": "Concept", "title_modern": "Persian
17    stratagem",

```

```

13   "description_original": "Persarum commentum
    irritum",
14   "description_modern": "A failed Persian military
    plan or ruse.", "evidence_spans": [1]}],
15   "relations": [
16     {"source": "Ammianus Marcellinus", "target": "
    Saracen customs", "relation_name": "describes",
17     "description": "Ammianus gives an ethnographic
    account of Saracen customs.",
18     "evidence_spans": [9, ..., 15]},
19     {"source": "Nohodares", "target": "Persian
    stratagem", "relation_name": "devised",
20     "description": "Nohodares planned the invasion
    of Batnae during its annual fair.",
21     "evidence_spans": [2, ..., 7]}]}

```

Extracted graph – Chinese chunk (Liu Mi, Sanjiao Pingxin Lun; romanized)

```

1  {"entities": [
2    {"type": "Person", "title_modern": "Liu Mi",
3     "description_modern": "Author of the Sanjiao
    Pingxin Lun.",
4     "evidence_spans": [25, 66]},
5    {"type": "Concept", "domain": "Religion &
    Spirituality", "title_modern": "fangbian (skillful
    means)",
6     "description_modern": "A Buddhist term for
    expedient teaching; here the expedient by which
7     earlier worthies reconciled the Three
    Teachings.", "evidence_spans": [2, 19, 20, 28]},
8    {"type": "Concept", "title_modern": "sanjiao (Three
    Teachings)",
9     "description_modern": "Confucianism, Buddhism,
    and Daoism; the text argues they can be unified.",
10    "evidence_spans": [25, 34]},
11    {"type": "Person", "title_modern": "Cheng Yi",
12    "description_modern": "Northern-Song Neo-
    Confucian; here decries Buddhism as heterodox and
    monstrous.",
13    "evidence_spans": [59, 90]},
14    "relations": [
15      {"source": "Liu Mi", "target": "sanjiao", "
    relation_name": "argues_for_unity",
16      "description": "Liu Mi holds the Three Teachings
    can be merged into one.", "evidence_spans":
    "[25, 26, 27]},
17      {"source": "Liu Mi", "target": "Cheng Yi", "
    relation_name": "refutes",
18      "description": "Liu Mi rebuts Cheng Yi, noting
    the Confucian classics also contain strange tales
    .",
19      "evidence_spans": [66, ..., 92]}]}

```

E.1 Corpus access and licensing

The source texts are under copyright and cannot be redistributed. The repository therefore ships catalogues for obtaining legally licensed copies together with derived, non-infringing artefacts. The Chinese corpus is built from *Zhongguo Xueshu Mingzhu Tiya* (*Compendium of Chinese Academic Masterworks*, Fudan University Press) (Zhou, 1992); the Latin corpus is built from *The Latin Library* (The Latin Library, 2026). For both, we release the work list and extracted graph and cluster data in JSONL, but no raw full text. For the peer-reviewed-paper benchmark, we provide the paper list and extracted research questions, findings, and cited primary-source evidence in JSON rather than copyrighted PDFs. Scripts reconstruct restricted components when redistribution is not allowed. Released data are limited to scholarly metadata, structured annotations, and derived representations needed for reproducibility.

F Prompt Contracts and Output Schemas

The released repository contains the literal system prompts, machine-readable schemas, and worked Chinese and Latin examples. This section records the contracts that determine each stage and the validation rules used in the experiments.

F.1 Agent workflow contracts

Discovering.

A research question becomes two to four sub-questions with languages, literal anchors, and retrieval granularity.

Annotating.

Passages return question-directed interpretations, links, and relevance in JSON keyed by source unit.

Comparing.

The two labeled sides of the shared EvidencePool return parallels, divergences, unilateral concepts, and inverted emphases. Analogues are treated as interpretive proposals, and both sides must supply evidence.

Referring.

Provenance checks produce follow-up queries naming a work, author, concept, or native term rather than a workflow instruction.

Sampling.

The EvidencePool yields 8–15 passage IDs with checks for question, source, and tradition coverage.

Illustrating.

Claims are exemplified with retrieved passages, and each verbatim quotation is bound to a stable [Eid].

Representing.

The shared synthesis call writes the essay, then emits a structured audit that preserves the same evidence IDs. Each audited quotation resolves to a retrieved source unit.

Reflection.

Coverage gaps and weakly supported claims may trigger up to two new retrieval rounds. Each round is tied to a diagnosed gap.

F.2 Graph, report, and audit schemas

Graph extraction returns `entities` and `relations`. Every entity contains a type from Concept, Person, or Work, a canonical title, a short description, and sentence-index evidence spans. Each directed relation names two extracted entities, their types, a relation label, a grounded description,

and its evidence spans. Relation endpoints must resolve to entities in the same validated output.

Community and semantic-cluster reports share one schema. It contains a title, summary, narrative, key findings, top entities, themes, a rating, and a rating explanation. Community reports summarize supplied graph relations. Cluster reports describe semantic affinity and mark it as distinct from documented textual connection. This distinction prevents embedding proximity from being reported as historical influence.

All structured stages pass JSON-schema validation. Source identifiers remain stable across retrieval, annotation, curation, prose citation, and the final audit. Cross-tradition questions require evidence from both source pools. The final audit records findings, evidence IDs, source coverage, and targeted follow-up queries. Ablations remove the corresponding stage while preserving the contracts of all remaining stages.

G Scholarly-Judge Rubric and Prompt

Both LLM raters and the human experts use one rubric, given verbatim below. In the blind comparative protocol the rater sees all six systems' anonymised essays for one question in random order and scores each independently on the same scale; it never sees the paper's evidence in the *free* track.

Scholarly-judge system prompt (humanities, comparative)	
1	You are a humanities scholar reviewing research-assistance material (not a publication-ready
2	paper). You are given the research question, the paper title that produced it (background only,
3	NOT a thesis to replicate), and the essay under review. In the comparative protocol you see
4	SEVERAL anonymised candidate essays for the SAME question in random order; score EACH
5	independently on the same 1-5 rubric. Length is not rewarded.
6	
7	1. AnswerAccuracy -- does the essay precisely address the relation the
8	question asks (how / why / in what sense / through what mechanism)
9	with textually sound claims, not merely list related concepts?
10	1 = off-topic / contradicts the texts;
11	2 = misreads the core concept; 3 = relevant but generic or half-answers the key relation
12	(topical adequacy CAPS HERE); 4 = accurate, minor simplification;
13	5 = precise on the core relation, comparanda kept distinct.
14	
15	2. ArgumentDepth -- is the argument driven by specific textual detail

```

16 and developed in layers, not
paraphrastic summary?
17 1 = no real argument; 2 = assertion/
paraphrase only; 3 = a thesis
18 but mainly generalisation or
restatement (SUMMARY CAPS HERE);
19 4 = structured multi-step argument,
several moves text-driven;
20 5 = layered analysis deriving claims
from close reading of specific
21 passages.
22
23 3. CoverageCompleteness -- does it cover
the sub-tasks the question
24 entails? (A comparison requires both
sides + similarity + difference
25 + the limits of the comparison.)
26 1 = misses major dimensions; 2 = one
side only; 3 = main question
27 answered but a key sub-dimension thin (
a single general point CAPS
28 HERE); 4 = main dimensions covered,
minor omissions; 5 = thorough
29 and handles nuance / limitations.
30
31 4. EvidenceQuality -- how well is the
answer grounded in specific
32 primary-text evidence: density of
direct quotation, specificity of
33 citation, close reading of the source
material, versus vague
34 work-name dropping or hand-waving
paraphrase? (Faithfulness floor:
35 treat a citation as fabricated only if
the quoted words clearly do
36 not belong to the cited work or are mis
-attributed; this is rare and
37 not the main concern.)
38 1 = essentially no primary-text
evidence or clear fabrication;
39 2 = evidence very sparse, loosely
connected; 3 = some real evidence
40 but sparse/generic or bare work-name
mention (~no direct quotation
41 CAPS HERE); 4 = key claims backed by
specific, accurate quotation;
42 5 = quotation specific and tightly
integrated with close-reading
43 analysis throughout.
44
45 Output VALID JSON only. Single essay:
46 {"AnswerAccuracy": <1-5>, "ArgumentDepth":
<1-5>,
47 "CoverageCompleteness": <1-5>, "
EvidenceQuality": <1-5>,
48 "rationale": "<one short sentence per
axis, in order>"}
49 Comparative (N essays):
50 {"scores": {"Answer 1": {"AnswerAccuracy":
<1-5>, ...},
51 "Answer 2": {...}, ...},
52 "comparative_note": "<which answer is
most grounded in direct
53 primary-text evidence, which is mostly
generic summary>"}

```

H Fragment-Level Qualitative Comparisons

Reading guide. The examples compare representative fragments from the same question. Green marks direct evidence, cyan marks evidence-to-claim reasoning, blue marks broad theses, orange marks source locators, purple marks graph summaries, and red marks limitations. Ellipses omit surrounding text.

H.1 Chinese Case Study

Question. 程颢与程颐的哲学思想有哪些相同与不同？这种异同分别体现在哪些方面，又有何社会根源与思想基础？

Fragment 1: Core distinction

Dimension: argument specificity

Naive LLM

程颢更强调“心即理”，主张“仁者浑然与物同体”；程颐则严格区分“性”与“心”……程颢的工夫更近于直觉体悟，程颐则更重理性分析。

Text RAG

程颢强调“识仁”与“诚敬存之”。其《识仁篇》云：“学者须先识仁。仁者，浑然与物同体……不须防检，不须穷索。”……程颐则更重“格物穷理”，提出“涵养须用敬，进学在致知”。

GraphRAG

本体论与工夫论的侧重不同：程颢更强调“仁”的感通与浑然一体；程颐更强调“理”的客观性和通过格物致知进行认知。……数据中“程颐主张物我一理”等关系指向了程颐这种主知、外向的工夫论特点。

SPIRE

二程本体论的最大分野，在于对“理”与“气”、“体”与“用”关系的不同理解。……程颐《周易程氏传》开篇即言：“易，变易也，随时变易以从道。”这句话虽简短，却蕴含了“理”与“气”、“体”与“用”的区分。……此种“分别而后统一”的思维方式，与程颢“浑然一体”的直接统一，形成鲜明对照。

Analysis. The Naive LLM states a plausible text-book contrast, Text RAG adds relevant quotations, and GraphRAG broadens the conceptual map. **SPIRE** turns a textual phrase into an interpretive mechanism: the contrast is derived from how the two thinkers handle *li*, *qi*, *ti*, and *yong*.

Fragment 2: Evidence use

Dimension: evidence-to-claim reasoning

Naive LLM

二程均重视格物致知，认为通过穷究事物之理可达天理。程颢强调“格物者，格至也”，程颐则系统阐述“格物穷理”的工夫。

Text RAG

后世学者概括：“明道以静字稍偏，不若专主于敬……伊川则以敬字未尽，益之以穷理之说。”……这说明二程在为学进路与工夫论侧重上有所不同。

GraphRAG

程颐提出“物我一理，明彼即晓此”……其工夫路径更侧重于“主敬”和“格物穷理”，强调一种严肃、有序的修养方法。

SPIRE

程颐的修养工夫以“识仁”为核心。门人问为学之道，答以“识得此理，以诚敬存之而已，不须防检，不须穷索”。……程颐则主“持敬”与“格物穷理”并进。他主张“涵养须用敬，进学则在致知”，认为道德的完成既需要内心的敬畏持守，也需要通过格物致知来穷尽事物之理。

Analysis. The systems differ in how they use citations. **SPIRE** uses paired quotations to construct a comparative argument about cultivation. Cheng Hao’s method is preservative and non-searching, while Cheng Yi combines reverent self-cultivation with inquiry into principle.

H.2 Latin Case Study

Question. How does Cicero’s *De Divinatione* compare evidence and explanation in assessing divination, and what does Marcus’ critique really depend on?

Fragment 1: What Marcus’ critique depends on

Dimension: claim precision

Naive LLM

Marcus’ critique depends on a distinction between correlation and causation: even if some predictions come true, this does not prove a causal link between the sign and the event. Marcus employs a form of skeptical reasoning. ...

Text RAG

Marcus’ critique, as presented in Book II, does not directly appear in the provided passages. ... The passages show that Quintus’ case is built on empirical observation and metaphysical commitment. Marcus’ likely strategy is to challenge the reliability of the evidence and the coherence of the explanation.

GraphRAG

Marcus' critique ultimately depends not on a dogmatic rejection of the gods, but on a rationalist and skeptical demand for a coherent, logical, and non-superstitious basis for belief.

SPIRE

Marcus' critique depends, in the final analysis, on a rigorous distinction between correlation and causation that Quintus systematically blurs. When Quintus presents the case of Crassus' disaster, ... he insists that the signs *non causas adferunt, cur quid eveniat, sed nuntiant evenitura* (I.30) [E5]. The distinction is revealing: Quintus himself admits that signs do not cause events, they only announce them.

Analysis. All four systems identify the same general issue, but they differ in how strongly the claim is grounded. **SPIRE** reconstructs the argument from a specific Latin sentence. It shows that the distinction between correlation and causation arises from Quintus' own formulation.

Fragment 2: Evidence versus explanation
Dimension: close reading

Naive LLM

Quintus marshals historical examples, dreams, oracles, and portents. ... Marcus argues that alleged successes are either coincidental, exaggerated, or self-fulfilling prophecies.

Text RAG

Quintus opens by asserting that divination is *the foretelling and foreknowledge of things that are thought to be fortuitous*. ... He compares the unfolding of time to the unwinding of a rope. ... The evidence gaps are significant: we lack Marcus' actual counterarguments.

GraphRAG

The critique demands a higher standard of evidence. Cicero contrasts the unreliable predictions of divination with reliable knowledge derived from other human arts, like medicine. ... Failures and ambiguities outweigh occasional successes.

SPIRE

Quintus observes that the ancients accepted divination *rerum magis eventis moniti quam ratione docti* (I.5) [E1], warned by outcomes rather than taught by reason. The phrase captures the tension that will dominate the entire debate: events provide the raw material, but reason must decide whether those events warrant the conclusions drawn from them.

Analysis. The baselines discuss evidence and explanation mostly as abstract categories. **SPIRE** identifies a compact Latin phrase and turns it into the organizing problem of the dialogue.

Fragment 3: Reconstructing the positive case
Dimension: balanced coverage

Naive LLM

Quintus relies on Stoic *sympatheia* and divine providence as explanatory frameworks, claiming that the gods communicate through signs that trained interpreters can read.

Text RAG

Quintus' explanation ties divination to a rational, causal framework: if all things happen by fate, and if the chain of causes can be grasped, then future events are in principle knowable through their signs.

GraphRAG

On one hand, Cicero argues for the validity of divination, grounding it in the existence of gods and their providential care for humanity, supported by logical arguments, appeals to fate and nature, and extensive historical examples.

SPIRE

Quintus writes: *Similis est haruspicum responsio omnisque opinabilis divinatio; coniectura enim nititur*. ... (II.24–25) [E2]. The logic is inductive: repeated observation of correlations between antecedent signs and subsequent events has produced an art. This comparison is crucial because it frames divination as a species of empirical knowledge. It is *opinabilis*, a matter of conjecture, but conjecture disciplined by experience.

Analysis. **SPIRE** first reconstructs Quintus' strongest case: divination as an empirical and conjectural art. This balanced account sharpens the later critique.

H.3 Agentic Baselines on a Matched Question

Question. How do Mencius and Xunzi differ in their understanding of *li* as a virtue, and what follows for the relation between moral and civic virtue?

This example comes from GLOBAL_ZH_007. Both agents retrieve from the shared store, so the comparison focuses on how they turn retrieved material into an answer.

PRG

For Mencius, *li* is an internal virtue rooted in the innate goodness of human nature, expressed through the four sprouts. ... For Xunzi, *li* is an external instrument created by sages to restrain and reform human nature. ... The result is an integrative model in Mencius, where moral and civic virtue are continuous and internally grounded, and a constructivist model in Xunzi, where civic virtue precedes and shapes moral virtue through external ritual discipline.

ReAct

As Mencius states: The heart of compassion is the sprout, *duan*, of *ren* ... the heart of reverence and respect, *gongjing zhi xin*, is the sprout of *li* ... I originally possess them (*Mencius* 6A6).

Analysis. PRG gives a clear and defensible contrast without a quotation or textual locator. ReAct retains the source terms and cites a specific passage, making its evidence use easier to inspect. Across the five categories, both agents remain below **SPIRE** in WorkR and SectionR (Table 12). The gap reflects the conversion of retrieved material into claims bound to specific evidence.