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

Anonymous ACL submission

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 a different mode of reasoning: interpretive, evidence-grounded argument over primary sources, where scholarly value depends on faithful quotation, verifiable provenance, and close reading. Existing research agents remain largely optimized for execution and retrieval, not evidence-grounded interpretive reasoning. To address this gap, we introduce **SPIRE** (**S**cholarly-**P**rimitives-**I**nspired **R**esearch **E**ngine), a multi-agent framework for evidence-grounded humanities scholarship. Drawing on Scholarly Primitives theory, **SPIRE** casts recurring humanities operations as cooperating agent roles (source discovery, evidence annotation, comparison, provenance checking, sampling, citation binding, and argumentative synthesis) over a multi-scale close-reading substrate of passages, intra-context graph communities, and cross-context semantic clusters. On a peer-reviewed-paper benchmark over classical Chinese and Greco-Roman Latin scholarship, **SPIRE** recovers cited primary-source evidence more reliably than Naive LLM, Text RAG, and GraphRAG, and receives higher blind-judge scores on answer accuracy, depth, coverage, and evidence quality. Ablations show that both the scholarly-operation agents and close-reading retrieval contribute to evidence-grounded essays. Code, data catalogues, and reproduction scripts are released anonymously at `https://anonymous.4open.science/r/SPIRE-CB4F`.

1 Introduction

LLM-based agents have recently achieved strong performance across scientific research domains. In **natural sciences** they design experiments, analyze data, and test hypotheses (Boiko et al., 2023; Lu et al., 2024; Ghafarollahi and Buehler, 2024; Romera-Paredes et al., 2024; Chen et al., 2025b); in **computer science** they survey literature, write and test code, and plan experiments (Schmidgall et al., 2025; He et al., 2025; Jimenez et al., 2024; Zhang et al., 2025; Zhuge et al., 2024); and in the **social sciences** they conduct quantitative analysis and simulate participants or populations (Park et al., 2023; Horton et al., 2023; Argyle et al., 2023; Li et al., 2023; Piao et al., 2025; Gao et al., 2024). These systems are enabled by 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., 2023c), chain-of-thought reasoning (Wei et al., 2022), reasoning–acting loops (Yao et al., 2023), self-reflection (Shinn et al., 2023), and reinforcement learning (Guo et al., 2025). Across these domains, however, reseach is typically organized around executable procedures, quantitative evaluation, or externally verifiable outcomes.

Humanities research requires a different kind of research assistance.[1] Unlike scientific workflows, it does not primarily require executable computation or reproduction of quantitative results. Its central work is interpretive and evidence-grounded: **close reading** of primary sources; **evidence mining and evidence-based reasoning**, which gather and weigh passages for and against a claim; and **original interpretation**, where argumentative logic itself is the contribution. It allows a wider range of well-argued readings, but still requires faithful quotation, verifiable provenance, and claims grounded in primary-source evidence. Fluent parametric answers are therefore insufficient when they fail to quote, contextualize, and reason from textual evi-

[1] By *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).

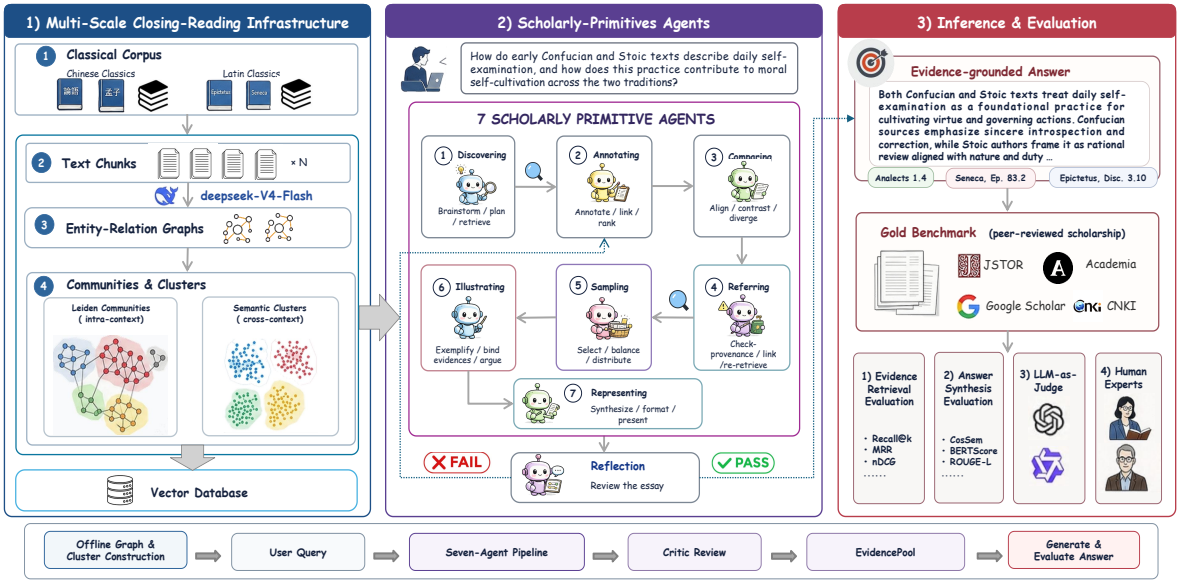

Figure 1: Overview of **SPIRE**: compiles classics into a multi-scale close-reading store, answers a research question with seven scholarly-primitive agents over an EvidencePool, and evaluates on a peer-reviewed-paper benchmark.

dence (Smith, 2016).

Humanities scholarship is interpretive and open-ended, yet still organized around recurring scholarly operations. Scholarly Primitives theory characterizes these operations as activities such as discovering, annotating, comparing, referring, sampling, illustrating, and representing (Unsworth, 2000). Building on this perspective, we operationalize recurring humanities research practices as executable multi-agent workflows: **SPIRE** (**S**cholarly-**P**rimitives-**I**nspired **R**esearch **E**ngine) is a multi-agent framework in which agents cooperate over a typed EvidencePool to discover sources, annotate evidence, compare sides, check provenance, sample passages, bind citations, and synthesize a citation-grounded argument. To our knowledge, **SPIRE** is among the first multi-agent research framework designed explicitly for evidence-grounded interpretive scholarship over primary sources, rather than isolated computational or digital humanities tasks.

Close reading depends on grounding interpretation in faithful primary-source evidence. RAG and its variants improve text search by embedding passages, and increasingly graphs or structured knowledge, for semantic retrieval (Lewis et al., 2020; Gao et al., 2023c; Edge et al., 2024; Luo et al., 2025; Han et al., 2025). However, humanistic reading requires more than retrieval: a scholar does not merely select relevant passages, but situates them in various contexts, 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.

Evaluation should therefore prioritize evidence-groundedness and defensible interpretation rather than the reproduction of a quantitative result. 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 a research question, a system must perform the scholarly steps and produce an evidence-grounded essay. We evaluate two complementary dimensions: cited-evidence recovery, testing whether the system retrieves primary sources used in the reference paper, and blind scholarly evaluation by both expert human reviewers and LLM judgesassessing answer accuracy, depth, coverage, and evidence quality. Compared with Naive LLM, Text RAG, and GraphRAG baselines, **SPIRE** attains 44.3% evidence recall versus $\leq 22.4\%$ for the strongest baseline (roughly double) and receives higher judge scores on all four aspects. Ablations show that downgrading individual curation primitives or close-reading retrieval tiers consistently reduces performance.

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), and are most developed 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 machine learning 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). Recent benchmarks such as ScienceAgentBench further formalize scientific discovery as executable workflows evaluated through code and execution (Chen et al., 2025b).

Humanities research poses a different target: its output is not an executed program, reproduced experiment, or simulated population, but an evidence-grounded interpretation of primary sources. The required operations are close reading, evidence selection, citation, comparison, provenance checking, and argumentative synthesis. Evaluations based on task completion, code execution, numerical reproduction, or generic NLG similarity therefore leave open how to design agents for interpretive scholarship, where success depends on evidence-grounded argument rather than executable verification.

2.2 Retrieval-augmented generation for research

Retrieval-augmented generation (RAG) grounds model outputs in retrieved passages to reduce unsupported claims and enable citation-backed answers (Lewis et al., 2020). For research, PaperQA answers over full-text articles (Lálá et al., 2023) and OpenScholar scales retrieval and self-feedback to millions of papers (Asai et al., 2024); graph-based variants add structured knowledge: GraphRAG via entity-relation graphs and community summaries, HyperGraphRAG via higher-order hyperedges (Edge et al., 2024; Luo et al., 2025). These have been applied across domains such as medicine (Wu et al., 2025; Kim et al., 2026) and technical documents (Knollmeyer et al., 2025).

But humanities evidence use differs from scientific literature survey: it asks not which paper supports an answer, but which primary-source passage, section, quotation, or summary bears on an interpretation. Standard RAG retrieves semantically similar passages and graph RAG expands to structural knowledge; neither models primary-source provenance as the basis of an interpretive argument.

Our evidence metric adapts *attributed text generation*, which scores how faithfully claims are supported by cited sources (Bohnet et al., 2022; Gao et al., 2023a,b), but specializes it to humanities provenance: matching at work, section, and sentence granularity, in the source language, against an open-ended interpretive target. Recent LLM work on digital humanities (intertextuality (Yang et al., 2025), historical-text analysis (Zeng, 2024), heritage retrieval (Kelly et al., 2025), text reuse (McManus et al., 2025; Duan, 2025), and classical translation and entity recognition (Chen et al., 2025a; Zhao et al., 2025; Xu et al., 2025), much of it at dedicated venues (Sprugnoli and Passarotti, 2024; Anderson et al., 2025; Hämäläinen et al., 2025)) each targets one task or corpus, not general citation-grounded assistance for primary-source research.

2.3 Scholarly Primitives and close-reading theory

The gaps above point back to humanities methodology itself: scholarly practice already provides operations for organizing interpretive work and close reading. Scholarly Primitives theory identifies common operations in humanities research: discovering, annotating, comparing, referring, sampling, illustrating, and representing (Unsworth, 2000). Work on scholarly information practices and humanities visualization similarly treats humanistic work as recurring practices of information seek-

ing, interpretation, organization, and representation (Palmer et al., 2009; Drucker, 2011). We use Scholarly Primitives because they summarize broad humanistic practice into repeatable operations that can be translated into agent steps.

Close-reading theory further explains the needed evidence substrate: interpretation does not only retrieve relevant passages but situates them within textual, historical, and conceptual relations. Contextual reading (Skinner and the Cambridge School) ties meaning to historically attested texts, authors, intentions, and arguments (Skinner, 1969); recurrent reading (Lovejoy's history of ideas) follows conceptual formations across periods, works, and traditions, including cases without direct textual links (Lovejoy, 1936). Though contextualists warn against decontextualized histories of ideas (Skinner, 1969; Armitage, 2012), the two modes answer different evidentiary needs. **SPIRE** therefore supports its Scholarly-Primitive agents with multi-scale close-reading 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** turns each scholarly primitive into an agent and each reading mode into a retrieval tier. A user supplies a classical-text corpus and an entity/relation schema, and **SPIRE** answers a research question with a evidence-grounded essay (Figure 1). *Seven scholarly-primitive agents* (§3.2) grow and curate a single shared EvidencePool from retrieval tiers, synthesize an argument, and re-ground it through reflection (≤ 2 rounds); they operate over a *multi-scale close-reading infrastructure* (§3.3) that compiles the corpus bottom-up into a persistent store of text chunks, an entity-relation graph, intra-context communities, and cross-context clusters. Evaluation centres on citation-grounded evidence retrieval and scholarly judging (§4).

Formally, let $\mathcal{C}$ be the corpus segmented into text units $u \in \mathcal{C}$. Given a research question $\mathcal{Q}$, **SPIRE** returns a ranked evidence pool $\mathcal{E}(\mathcal{Q}) = [u_1, \ldots, u_K]$ and a structured answer $A(\mathcal{Q})$ that factor as

$$P(A, \mathcal{E} \mid \mathcal{Q}) = \underbrace{P(\mathcal{E} \mid \mathcal{Q}, \mathcal{C})}_{\text{retrieve (§3.3)}} \underbrace{P(A \mid \mathcal{Q}, \mathcal{E})}_{\text{synthesize (§3.2)}}, \quad (1)$$

i.e. retrieve, then synthesize; humanities assistance is evaluated on both stages.

3.2 Scholarly-primitive agents

3.2.1 Primitives as typed agents

	Primitive	**SPIRE** role and execution
$\mathcal{D}$	Discovering	Plan sub-questions, anchors, hints; query all tiers.
$\mathcal{N}$	Annotating	Annotate and re-rank by interpretive significance.
$\mathcal{P}$	Comparing	Align sides; contrast parallels and divergences.
$\mathcal{R}$	Referring	Check provenance, link sources, flag absent works.
$\mathcal{S}$	Sampling	Select a quota-balanced, diverse evidence subset.
$\mathcal{I}$	Illustrating	Exemplify and bind claims to cited evidence.
$\mathcal{T}$	Representing	Synthesize bound evidence; format and present.

Table 1: The seven scholarly primitives, each realized as one agent role in **SPIRE**; the ablation (§5.3) removes one agent at a time.

Definition 1 (EvidencePool)**.** The *EvidencePool* $\mathcal{E}$ is the mutable state shared by all agents: an ordered set of evidence items, each a tuple $(u, t, s_w, s_a, s_{\cos}, \ell, \text{lang}, \sigma)$ recording a retrieved text unit u, its retrieval tier t, work-filter / anchor / cosine scores $(s_w, s_a, s_{\cos})$, canonical locator ℓ, language, and comparative side label σ. Each agent is typed by how it acts on $\mathcal{E}$: an *evidence-pool transformer* $a : \mathcal{E} \to \mathcal{E}'$ that grows or curates the pool, or a *synthesizer* $a : \mathcal{E} \to A$ that writes the argument from it.

SPIRE realizes the seven primitives (§2.3) as cooperating typed agents in a streaming workflow. In Unsworth's order they are Discovering ($\mathcal{D}$), Annotating ($\mathcal{N}$), Comparing ($\mathcal{P}$), Referring ($\mathcal{R}$), Sampling ($\mathcal{S}$), Illustrating ($\mathcal{I}$), and Representing ($\mathcal{T}$), in three roles (Table 1): research framing ($\mathcal{D}$); close reading and evidence curation ($\mathcal{N}, \mathcal{P}, \mathcal{R}, \mathcal{S}$); and writing the argument ($\mathcal{I}, \mathcal{T}$). By Definition 1, each is an *evidence-pool transformer* $a : \mathcal{E} \to \mathcal{E}'$ ($\mathcal{D}, \mathcal{N}, \mathcal{R}, \mathcal{S}$) or a *synthesizer* $a : \mathcal{E} \to A$ ($\mathcal{P}, \mathcal{I}, \mathcal{T}$) over the shared EvidencePool, which Discovering initializes, the curation agents refine, and the synthesizers consume.

SPIRE runs all seven agents on the LLM with up to two reflection rounds. To ablate a primitive, we remove its agent from the workflow one at a time, leaving its input pool unchanged (§5.3); per-agent prompts are in App. F.

3.2.2 Retrieval, orchestration and reflection

Given $\mathcal{Q}$, Discovering ($\mathcal{D}$) emits a *query plan* $\pi(\mathcal{Q}) = (\{q_i\}, \text{lang}, c, \text{anchors}, \text{hints})$: it anchors the works, authors, and language in $\mathcal{Q}$ against corpus metadata, sets the comparison type $c \in$

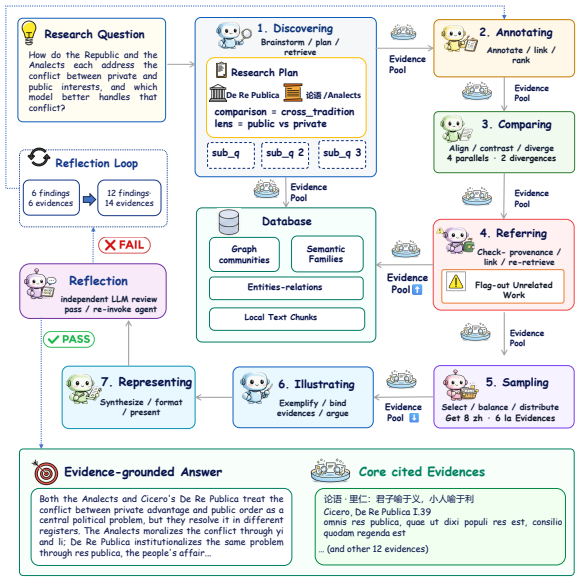

Figure 2: **SPIRE** on a concrete cross-tradition run (Cicero's *De Re Publica* vs. the *Analects*).

{none, intra-text, intra-author, intra-tradition, cross-tradition}, and decomposes $\mathcal{Q}$ into sub-questions q_i tagged by granularity and perspective (the hints broaden recall but license no unsupported claim). Each q is encoded once and scored by the BGE-M3 inner product $\mathrm{sim}(q, x) = \phi(q)^\top \phi(x)$ against three tools over the multi-scale store (§3.3): **local** (L, top chunks/entities, boosted when q names a tied work), **community** (intra-context, G, scoring community summaries expanded to member units), and **cluster** (cross-context, I, scoring semantic clusters expanded to member-entity units, surfacing evidence missed by graph edges). Each unit carries $(t, s_w, s_a, s_{\cos})$ (tier, work-filter, anchor, cosine); results form the shared EvidencePool, the ranked union over sub-questions and tools,

$$E(\mathcal{Q}) = \mathrm{rank}\Big(\bigcup_{i} \bigcup_{t\in\{L,G,I\}} T_t(q_i), \mathrm{score}\Big),$$
$$\mathrm{score}(u) = (t, s_w, s_a, s_{\cos}), \quad L \succ G \succ I. \tag{2}$$

sorted lexicographically, so an explicit work or anchor match outranks raw cosine: a conservative, provenance-first rule mirroring how humanists weight documentary evidence over surface similarity, rather than a single tunable weighted sum. The orchestrator then runs $\mathcal{D} \to \mathcal{N} \to \mathcal{P} \to \mathcal{R} \to \mathcal{S} \to (\mathcal{I}{+}\mathcal{T}) \to$ Reflection over this pool: $\mathcal{N}$ re-ranks by interpretive significance; $\mathcal{P}$ aligns the two sides; $\mathcal{R}$ checks provenance, links sources, and flags works named in $\mathcal{Q}$ but absent; $\mathcal{S}$ selects a representative subset under per-work, side, and sub-question quotas; and $\mathcal{I}{+}\mathcal{T}$ synthesizes, Illustrating binding claims to citations and emitting the structured evidence fields. Representing writing the final prose. A separate Reflection step (≤ 2 rounds) reviews the draft and, unless it passes, loops back to Annotating ($\mathcal{N}$) with an updated retrieval.

Figure 2 traces this on a cross-tradition question (Cicero's *De Re Publica* vs. the *Analects*; `zh+la`): the agents increasingly build a bilingual pool via local, community, and cluster search, curate and sample it, and synthesize citation-grounded findings from 14 Chinese and Latin evidences.

	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

3.3 Multi-scale close-reading infrastructure

3.3.1 Corpora and preprocessing

We instantiate **SPIRE** on two classical corpora central to humanities scholarship: classical Chinese (*Zhongguo Xueshu Mingzhu Tiyao* (Zhou, 1992)) and Greco-Roman Latin (*The Latin Library* (The Latin Library, 2026)). History graduate students annotate every work with metadata including author, date, alternative titles. Both span antiquity to the modern period and are standard primary sources in history, philology, and literary study (sizes and spans in Table 2). We segment each with `tiktoken` into 1000-token chunks (100-token overlap), giving units $\mathcal{U} = \{u_1, \ldots, u_M\}$. Graph extraction and community/cluster reports use DEEPSEEK-V4-FLASH (thinking; temperature 0.1; App. G), while the answer-generation agents use the DEEPSEEK-V4-FLASH generator setting in §4.2; one multilingual encoder, BGE-M3 (Chen et al., 2024a) (d=1024, dense), embeds every object and query, so a single encoded query scores against every tier and matches Chinese and Latin without a per-tier or per-language encoder.

3.3.2 The multi-scale store

A user defines an entity/relation schema and an LLM extracts a typed graph from the chunks. Our exemplar is *conceptual history* (shared across history, literary studies, and philology): three entity types (*Concept*, *Person*, *Work*) and free, text-grounded relations, with a few-shot prompt (one

Chinese, one Latin example; App. G). DEEPSEEK-V4-FLASH extracts; format validation, deduplication, and cross-chunk entity merging then yield a typed graph $\mathcal{G} = (\mathcal{V}, \mathcal{R})$ with $\mathcal{R} \subseteq \mathcal{V} \times \mathcal{V}$ directed semantic relations (per-tradition counts in Table 2). Another field's schema substitutes without pipeline changes (examples in App. E).

Two higher-level tiers sit on the graph, each operationalizing a complementary mode of close reading (§2.3). The **intra-context** tier captures corpus-attested relational context among works, persons, and concepts, following contextual reading (Skinner, 1969); the **cross-context** tier captures recurrent conceptual affinity among semantically akin entities lacking explicit graph edges, following history-of-ideas (Lovejoy, 1936; Betti and van den Berg, 2016). We build them with Leiden community detection (Traag et al., 2019) on $\mathcal{G}$ and HDBSCAN (McInnes et al., 2017) over entity-description embeddings, respectively, and DEEPSEEK-V4-FLASH names and summarizes each community/cluster from its top entities. Both are built per tradition at four fine-to-coarse resolutions, giving parallel multi-scope views (per-scale counts in App. Table 19).

A normalized relational schema maps each community/cluster to its member entities and relations, each entity and relation to its supporting chunks, and each text unit to its work, chapter, section, and canonical citation, bridging chunked text and scholarly reference. All objects are embedded by ϕ and stored in one DuckDB store with FAISS inner-product indexes (Johnson et al., 2019).

4 Experiments

4.1 Benchmark construction

We build a peer-reviewed-paper benchmark in which each paper supplies a humanist answer and the primary-source evidence it rests on. We collect 406 refereed papers (English or Chinese) on classical Chinese and Latin texts, single- and cross-tradition, from JSTOR, Academia, Google Scholar, and CNKI (Table 3). Two graduate students in history and philology read each paper and manually extract a record: research question, abstract, findings, and cited primary-source evidence, each anchored at a verbatim quotation, a section or chapter, or a work (example in App. E). We extract by hand rather than with an LLM because humanities papers have no fixed structure and mark primary evidence through formatting an automatic extractor

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).

misses (indented passages, regular-script Chinese, italicized Latin), so finding the first-hand evidence a claim rests on requires close reading. A system sees only the title and question and must return an evidence-grounded essay. We score it automatically against the paper and, since a humanities answer turns on its argument and use of evidence rather than a single reference, by two human and two LLM raters on rubrics blind to the paper's evidence.

4.2 Baselines and setup

All systems share the corpus and substrate of Table 2, the BGE-M3 encoder (Chen et al., 2024a) (d=1024), the DEEPSEEK-V4-FLASH generator (DeepSeek-AI, 2026) (temperature 0.3; a robustness re-run with a GEMINI-3-FLASH generator is reported in App. B), the 406-paper gold, and the same evaluation code. We compare **SPIRE** against a controlled ladder of baselines, each in its standard published configuration: **Naive LLM** answers from parametric memory with no retrieval; **Text RAG** is the standard retrieve-then-read setup (Lewis et al., 2020) with text-chunk retrieval; and **GraphRAG** is the released GraphRAG configuration (Edge et al., 2024), with entity-graph and community-summary retrieval over the same graph. We do not modify their retrieval logic, so any gap reflects **SPIRE**'s design rather than a re-tuned baseline. §5.3 ablates the curation primitives (contribution 1) and the retrieval tiers (contribution 2), isolating the seven-primitive decomposition from raw retrieval access. **SPIRE** runs the `full` seven-agent configuration throughout.

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.

System	eR	eR@10	workR@10	sentR@10	secR@10	MRR@10	nDCG@10
Naive LLM†	14.3	14.3	17.4	3.6	4.4	15.7	15.0
Text RAG	22.4	14.5	12.5	2.7	3.5	6.5	4.0
GraphRAG	12.4	12.5	13.2	2.3	3.6	6.2	8.9
SPIRE (full)	**44.3**	**24.0**	**42.4**	**5.6**	**15.3**	**33.5**	**28.9**

Table 4: Evidence retrieval at k=10 (paper-level macro-mean, %); best per column in **bold**. eR is rank-independent full-pool recall. †Naive LLM's from-memory citations are scored against their nearest corpus chunks.

	GPT-5.4				Qwen3.5				Human A				Human B			
System	Acc	Dep	Cov	EQ	Acc	Dep	Cov	EQ	Acc	Dep	Cov	EQ	Acc	Dep	Cov	EQ
Naive LLM	3.72	2.81	3.65	2.26	4.01	2.74	3.58	2.46	3.62	2.96	2.51	2.26	4.25	2.06	2.24	1.47
Text RAG	1.78	1.46	1.57	1.55	1.65	1.25	1.45	1.30	1.73	1.17	1.00	1.30	1.76	1.00	1.00	1.00
GraphRAG	2.35	2.37	2.61	1.68	2.55	1.93	2.50	1.39	2.30	2.08	1.85	1.43	2.66	1.40	1.66	1.00
SPIRE (full)	**4.46**	**4.63**	**4.76**	**4.50**	**4.95**	**4.96**	**4.96**	**4.90**	**4.46**	**3.89**	**4.39**	**3.99**	**4.94**	**4.57**	**4.62**	**4.76**

Table 5: Scholarly judge scores (1–5), *free* track; best per rater in **bold**. Acc/Dep/Cov/EQ: AnswerAccuracy, ArgumentDepth, CoverageCompleteness, EvidenceQuality. Inter-rater agreement (Gwet's AC2) in App. A.5.

Gold evidence is considered recovered when its source work is retrieved and any of three signals fires: a verbatim/paper-context substring match, a section/chapter-locator match, or a BGE-M3 cosine above 0.8 within-language (0.7 across, for translated quotations). We report recall at three provenance granularities: sentR (exact sentence), secR (cited section), and workR (source work); plus rank-independent eR over the full pool and MRR/nDCG (Järvelin and Kekäläinen, 2002) on secR. Recalls are macro-averaged across papers (%); Table 4 reports k=10, the full $k \in \{5, 10, 20\}$ sweep App. Table 7.

Answer similarity. As a standard text-generation measure, we also report each essay's surface similarity to the paper's abstract and findings (BGE-M3 cosine, BERTScore (Zhang et al., 2020), ROUGE-L (Lin, 2004); full numbers in App. A).

Scholarly judge. Our *primary* arbiter is human: two humanities scholars (classics; philosophy) score each answer 1–5 on AnswerAccuracy, ArgumentDepth, CoverageCompleteness, and EvidenceQuality (rubric in App. H), over a stratified sample of 100 papers (20 per category of Table 3). The protocol is *free* and comparative: raters see only the question and the four systems' anonymised, shuffled answers, never the reference paper's own answer. Two LLM raters (GPT-5.4, QWEN3.5-122B-A10B) replicate this as a scalable proxy (Liu et al., 2023; Zheng et al., 2023; Shen et al., 2026); they also run a *reference* track (additionally shown the paper's findings), which preserves the ranking (App. H).

Cost. Per-paper latency, token usage, and agent-step count (App. Table 14).

5 Results and Discussion

5.1 Main results

Table 4 reports automatic evidence and answer metrics, and Table 5 reports the per-aspect scholarly judge, comparing **SPIRE** with the baselines. Table 6 isolates the contribution of agents and retrieval tiers through ablations.

5.2 Analysis and discussion

On grounded evidence **SPIRE** leads every baseline over the full 406-paper benchmark: eR=44.3% vs. ≤22.4% for the strongest baseline (Table 4), roughly double, and the margin widens on provenance-precise recovery (workR=42.4 vs. ≤17.4; secR=15.3 vs. ≤4.4; sentR=5.6 vs. ≤3.6; MRR=33.5 vs. ≤15.7). GraphRAG recovers the *fewest* primary passages (eR=12.4, below Text RAG's 22.4) because community-summary retrieval abstracts the citation-bearing passage away, whereas **SPIRE** keeps local passages primary and fuses community and cluster context by Eq. 2 (§2.2). Crucially, the gain is provenance, not surface reach: answer-to-reference similarity does *not* track grounding, with the no-retrieval Naive LLM close to **SPIRE** on findings cosine (0.810 vs. 0.818) and above it on BERTScore (85.3 vs. 84.4), while the terse Text RAG is lowest (0.654 on findings; App. Table 7); a fluent from-memory essay can therefore resemble the paper as much as a grounded one. **SPIRE** recovers substantially more citation-bearing work- and section-level evidence;

this measures evidence alignment, not interpretive uniqueness.

SPIRE leads the scholarly judge on every aspect, for both LLM raters and both human experts (Table 5), and raters prefer it to each baseline on at least 96 of the 100 judged papers. The pattern is diagnostic. The no-retrieval Naive LLM stays on-question (AnswerAccuracy ≈3.9) yet cannot ground its claims (EvidenceQuality ≈2.1): it names works without quoting them. Text RAG and GraphRAG do retrieve, but quote tersely, only in the paper's language, or refuse, scoring low on *every* aspect (AnswerAccuracy ≤2.7, EvidenceQuality ≤1.7). Only **SPIRE** is high throughout (≥3.9 on every aspect across all four raters), grounding claims in original-language quotation with translation plus later commentary; its margin is widest on EvidenceQuality and ArgumentDepth (fragment-level comparison, App. I). Both experts also flagged that the Text RAG and GraphRAG baselines render primary sources only in the paper's language, and that graph retrieval can reach a secondary commentator rather than the primary text. On reliability we report Gwet's AC2 (quadratic weights) throughout, because per-aspect weighted κ collapses under the top-heavy score marginals here: for example, on **SPIRE** the human raters agree at AC2 0.88 yet weighted κ falls to 0.14, the kappa paradox. The two LLM raters agree strongly (AC2 0.88–0.95), the two human experts at AC2 0.79–0.88, and human–LLM agreement on **SPIRE** and Naive LLM (the two systems spanning the score range) is high (AC2 0.80–0.91), supporting the LLM judge as a proxy (App. A.5). As LLM-driven agents, **SPIRE** inherits LLM limits on questions lacking a clear scope (App. I), assisting scholars rather than replacing them.

5.3 Ablation studies

We ablate each primitive (removing its agent) and each retrieval tier, holding corpus, substrate, encoder, and generator fixed (Table 6); only Discovering (the sole evidence input) and the Representing writer cannot be removed. Every component helps, with each removal lowering both evidence recovery and the judge below **SPIRE**-full (eR 44.3, judge 3.44/4.01), along two lines. *(i) The multi-scale substrate drives grounding* (contribution 2): disabling either tier cuts eR to ∼32–33, and to 28.5 with both removed (local only), with secR and the other recall and ranking metrics falling in step (App. Table 13). *(ii) Curation shapes argument quality more than raw recall* (contribution 1): removing any of $\mathcal{N}/\mathcal{P}/\mathcal{R}/\mathcal{S}$ moves eR only 2–5 points yet steadily lowers the LLM judge (GPT-5.4 3.44 → 3.06–3.18), most for Referring and Comparing. Illustrating is the single most consequential primitive, with the lowest eR (27.8, below even local-only) and lowest LLM-judge score (3.07/3.74), because its claims drive **SPIRE**'s reflective second retrieval pass, so removing it degrades discovery, not just synthesis. **SPIRE**'s grounding gains come at ∼3× Text RAG's latency (App. Table 14).

Configuration	eR	workR	sentR	secR	GPT*	Qwen*
SPIRE (full, 7 primitives)	44.3	42.4	5.6	15.3	3.44	4.01
w/o one agent:						
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
Retrieval tier:						
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 6: Component ablations of **SPIRE** (k=10, paper-level macro-mean %). Agents $\mathcal{N}, \mathcal{P}, \mathcal{R}, \mathcal{S}, \mathcal{I}$ are as in Table 1; tier rows disable a retrieval tier. *GPT/Qwen are LLM-judge overall scores (mean of the four aspects, 1–5). MRR, nDCG, the $k \in \{5, 10, 20\}$ sweep, and answer similarity are in App. Table 13.

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, and GraphRAG, 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

Instantiated traditions. We instantiate and evaluate **SPIRE** on classical Chinese and Greco-Roman Latin corpora, two rich and demanding reservoirs of humanistic scholarship. These traditions provide strong testbeds for evidence-grounded reasoning because they require multilingual source recovery, close reading, commentary awareness, and cross-text comparison. Future work will extend the same framework to a broader range of humanities domains, including manuscript studies, vernacular and modern literatures, religious studies, archival history, and scholarship in additional languages and historiographic traditions.

Textual evidence as foundation. This paper focuses on textual evidence—passages, works, chapters, quotations, and relations among texts—because such materials are foundational to classical studies, philology, philosophy, history, literature, and many adjacent fields. As humanities research increasingly brings together texts with images, archaeological objects, performance records, maps, and archival materiality, an important next step is to extend **SPIRE** with multimodal retrieval, provenance tracking, and agent analysis over non-textual evidence while preserving the same emphasis on verifiable grounding.

Human-centered scholarship. **SPIRE** is designed to assist humanistic inquiry by making source discovery, citation checking, comparison, and evidence-bound synthesis more inspectable and easier to teach. This is especially useful for students and early-stage researchers learning how to move from sources to arguments, and for experts who want a transparent assistant for exploring large corpora. Humanistic interpretation remains scholar-led: the framework is intended to support close reading, pedagogy, and peer review, while keeping final interpretive authority with researchers.

Ethics Statement

We evaluate AI assistance for humanities research rather than replacement of humanities researchers. **SPIRE** supports evidence discovery, citation checking, and interpretive drafting in a provenance-aware workflow, with the goal of widening access to rigorous source-based scholarship and helping students, early-stage researchers, and experts inspect the evidentiary path from primary texts to arguments. The framework should be used as a human-centered research assistant: it strengthens documentation and review, while scholarly judgement, teaching responsibility, and peer evaluation remain with human researchers.

Artifact use and data protection. The released artifacts are intended for non-commercial research and evaluation of evidence-grounded humanities research assistants. They should not be used to replace scholarly judgment, generate uncited claims, or redistribute restricted source materials outside their original access conditions. The benchmark is built from published scholarship and classical primary sources rather than private personal data. Author identities from paper-derived records are not used as model inputs or evaluation targets, and released benchmark records are anonymized or restricted according to source access conditions.

Human annotation and evaluation. Annotators and expert raters were recruited through academic collaboration rather than crowdsourcing. They were graduate students or scholars with relevant expertise in history, philology, classics, or philosophy. They were compensated or credited according to local institutional practice, and no demographic information was collected. All annotators and raters were informed that their annotations and ratings would be used for research evaluation and reported only in aggregate or anonymized form.

Code and Data Release

Code, data store, and benchmark are released anonymously at `https://anonymous.4open.science/r/SPIRE-CB4F`. We release the code under the MIT License and the benchmark metadata under CC BY-NC 4.0 where redistribution is permitted. The source texts are under copyright and we cannot redistribute them, so the repository instead ships the catalogues needed to obtain legally licensed copies, together with our derived, non-infringing artefacts. The Chinese corpus is built from *Zhongguo Xueshu Mingzhu Tiyao* (*Compendium of Chinese Academic Masterworks*, Fudan University Press) (Zhou, 1992); the Latin corpus from *The Latin Library* (The Latin Library, 2026). For both, we release a plain-text bibliography (the work list) so that others can locate and download legally licensed editions, alongside the extracted graph and cluster data in JSONL (no raw full text). For the peer-reviewed-paper benchmark (humanities articles that are typically absent from

open preprint servers and often require purchase or institutional access), we likewise provide the paper list and the extracted benchmark (research questions, findings, and cited primary-source evidence, in JSON) in JSON rather than the PDFs. Source texts and paper-derived records are used and distributed only according to their original access conditions; scripts are provided to reconstruct restricted components when redistribution is not allowed. All released data is limited to scholarly metadata, structured annotations, and derived representations needed for reproducibility, and avoids personal or sensitive information.

References

Adam Anderson, Shai Gordin, Bin Li, Yudong Liu, Marco C. Passarotti, and Rachele Sprugnoli, editors. 2025. *Proceedings of the Second Workshop on Ancient Language Processing*. 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.

David Armitage. 2012. What's the big idea? intellectual history and the longue durée. *History of European Ideas*, 38(4):493–507.

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)*.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, and 3 others. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

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.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2025a. Benchmarking LLMs for translating classical Chinese poetry: Evaluating adequacy, fluency, and elegance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33019–33036, Suzhou, China. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335. 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. In *arXiv 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. 2025b. 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`. Model card for DeepSeek-V4-Flash.

Johanna Drucker. 2011. Humanities approaches to graphical display. *Digital Humanities Quarterly*, 5(1).

Siyu Duan. 2025. Quantitative intertextuality from the digital humanities perspective: A survey. *arXiv preprint arXiv:2510.27045*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv 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.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023c. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Alireza Ghafarollahi and Markus J. Buehler. 2024. SciAgents: Automating scientific discovery through multi-agent intelligent graph reasoning. *arXiv preprint arXiv:2409.05556*.

Daya Guo, Dejian Yang, Haowei Zhang, and 1 others. 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.

Mika Hämäläinen, Emily Öhman, Yuri Bizzoni, So Miyagawa, and Khalid Alnajjar, editors. 2025. *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*. Association for Computational Linguistics.

Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A. Rossi, Subhrabrata 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). *arXiv 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.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

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 Caymazer, 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 Lálá, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, 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.

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.

Stuart M. McManus, Yuji Li, Leo Tam, and 1 others. 2025. Measuring accommodation: A quantitative text-reuse analysis of matteo ricci's reliance on Chinese classics. *International Journal of Digital Humanities*.

Franco Moretti. 2013. *Distant Reading*. Verso.

Carole L. Palmer, Lauren C. Teffeau, and Carrie M. Pirmann. 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 simulacra 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, Xinchi Qiu, Chenxi Whitehouse, Lisa Alazraki, and 1 others. 2026. Rethinking rubric generation for improving LLM judge and reward modeling for open-ended tasks. *arXiv 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.

Rachele Sprugnoli and Marco Passarotti, editors. 2024. *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*. ELRA Language Resource Association.

The Latin Library. 2026. The Latin Library. `https://www.thelatinlibrary.com/`. Accessed 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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

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*.

Qiankun Xu, Yutong Liu, Dongbo Wang, and 1 others. 2025. Automatic recognition of cross-language classic entities based on large language models. *npj Heritage Science*.

Yue Yang, Yinzhi Xu, Chenghao Huang, JohnMichael Jurgensen, Han Hu, and Hao Wang. 2025. InterIDEAS: Philosophical intertextuality via LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23151–23172, Suzhou, China. Association for Computational Linguistics.

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*.

Yifan Zeng. 2024. HistoLens: An LLM-powered framework for multi-layered analysis of historical texts — a case application of Yantie Lun. *arXiv preprint arXiv:2411.09978*.

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*.

Zhixiao Zhao, Guangyao Sun, Chang Liu, and 1 others. 2025. Research on machine translation of ancient books in the era of large language model. *npj Heritage Science*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao 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, Zheye 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.

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

A.1 Evidence-recall matching rule

A gold evidence is recovered when its source work is retrieved and a verbatim or paper-context substring matches (after NFKC normalization and punctuation stripping), a chapter locator matches, or the BGE-M3 cosine clears a threshold (τ_s=0.80 same-language, τ_c=0.70 cross-language, since some papers quote their sources in translation); aggregation is paper-level macro-mean. Sensitivity to τ_s, τ_c is reported in §A.4.

A.2 Retrieval and answer-similarity sweep ($k \in \{5, 10, 20\}$)

Table 7 gives the full $k \in \{5, 10, 20\}$ retrieval / answer sweep (behind the k=10 Table 4); Table 14 gives the runtime cost. The answer-similarity columns (cosSem, R-L, BScore) appear only here: they are close across all systems and do not separate grounded from ungrounded answers, confirming that surface similarity to the reference paper is a weak measure of grounding quality and belongs outside the main results (§4.3).

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

Table 8 breaks the 406-paper retrieval and answer-similarity results down by the five benchmark categories of Table 3 (k=10, macro-mean over applicable papers). The headline pattern is uniform:

System	k	eR	eR@k	workR	sentR	secR	MRR	nDCG	cosSem	R-L	BScore
Naive LLM†	5	–	13.9	16.8	3.4	4.2	15.6	11.4	–	–	–
	10	14.3	14.3	17.4	3.6	4.4	15.7	15.0	81.0	9.0	85.3
	20	–	14.3	17.4	3.6	4.4	15.7	17.6	–	–	–
Text RAG	5	–	10.6	9.2	2.1	2.5	5.9	3.8	–	–	–
	10	22.4	14.5	12.5	2.7	3.5	6.5	4.0	65.4	3.4	83.8
	20	–	17.5	16.1	3.5	4.5	6.8	4.4	–	–	–
GraphRAG	5	–	11.8	12.3	2.2	3.2	5.8	5.3	–	–	–
	10	12.4	12.5	13.2	2.3	3.6	6.2	8.9	79.0	9.9	84.0
	20	–	12.4	13.2	2.3	3.6	6.2	11.1	–	–	–
SPIRE (full)	5	–	19.3	38.5	4.1	10.1	32.7	28.0	–	–	–
	10	**44.3**	**24.0**	**42.4**	**5.6**	**15.3**	**33.5**	**28.9**	81.8	10.9	84.4
	20	–	28.3	45.1	7.8	18.6	33.8	30.3	–	–	–

Table 7: Full retrieval / answer sweep over the 406-paper benchmark, $k \in \{5, 10, 20\}$ (paper-level macro-mean, %). eR (full-pool, rank-independent) and the answer-similarity columns cosSem / R-L / BScore (each essay vs. the paper findings) carry no k and appear on the k=10 row only, with "–" elsewhere. Best retrieval metric at k=10 in **bold**. The answer-similarity columns do *not* track grounding: the no-retrieval Naive LLM is close to **SPIRE** on cosSem and higher on BScore while the terse Text RAG scores lowest, and these columns are left unbolded. †Naive LLM has no retrieval; its from-memory citations are resolved to their nearest corpus chunks for scoring.

SPIRE leads full-pool evidence recall (eR) in *every* category (36.5–52.6 vs. $\leq$30.1 for any baseline), and the lead on work-level recovery (workR) and ranking (MRR) is at least as large. Two structure-dependent effects stand out. (i) *Flat retrieval collapses on cross-tradition questions*: Text RAG, GraphRAG, and Naive LLM all fall to $\leq$13.8 eR on CROSS (and $\leq$7.4 workR), because pure similarity retrieval cannot balance the two traditions and drifts to one side, whereas **SPIRE**'s language-balanced, work-anchored retrieval holds 36.5. (ii) *The gain is widest on Chinese single/multi-text close reading* (LOCAL_ZH, GLOBAL_ZH: eR 52.6/43.7 vs. Text RAG 30.1/18.9), where section- and work-level provenance (secR up to 21.2 vs. $\leq$1.3) separates the systems most sharply. Answer similarity (findings cosine) tells the opposite story within every stratum: it is high and nearly flat across systems (0.63–0.83), with the no-retrieval Naive LLM matching or exceeding **SPIRE** in three categories—reconfirming, per category, that surface similarity does not measure grounding.

A.4 Threshold sensitivity

The evidence matcher in the main results uses $(\tau_s, \tau_c) = (0.80, 0.70)$ (§4.3; §A), fixed *a priori* as conventional BGE-M3 semantic-match operating points; they were *not* tuned on a labelled development split. We deliberately avoid fitting the matcher to our own benchmark, because these thresholds govern the *evaluation* (when a retrieved passage counts as recovering a gold evidence) rather than any system's behaviour. The headline scores are reported in Table 4 and Table 7. As an additional diagnostic audit of threshold robustness, Table 9 sweeps both thresholds from very permissive $(0.70, 0.60)$ to very strict $(0.90, 0.80)$ for **SPIRE** (full), and Table 10 reports the same diagnostic full-pool evidence recall (eR) comparison for every system.

Three patterns hold across the sweep. (i) Work-level recall is invariant (workR@10 $= 51.2$ at every setting), because it depends only on whether a gold evidence's source work enters the top-k, not on sentence-level cosine. (ii) Sentence recall is the most threshold-sensitive metric (sentR@10 falls from 18.2 to 12.2), being the signal most reliant on semantic matching; section recall moves less ($23.8 \rightarrow 21.8$) because it is also supported by locator and chapter matches, and aggregate evidence recall declines smoothly ($57.9 \rightarrow 46.5$) as semantic-only matches are filtered out. (iii) Most importantly, the *ranking of systems is stable at every threshold*: **SPIRE** leads on evidence recall under all five settings (Table 10), by $\sim$18 points over the strongest baseline. The main-paper setting $(0.80, 0.70)$ sits in the middle of this range, so it neither flatters nor penalises any system relative to the alternatives. These tables are a diagnostic threshold audit rather than a replacement for the headline retrieval table; Table 4 and Table 7 remain the source of the main reported scores.

A.5 Inter-rater agreement: AC2 vs. weighted κ

We report two complementary chance-corrected agreement coefficients on the *free*-track scores, both with quadratic (ordinal) weights: quadratic-

Category	System	eR	workR	secR	MRR	sSemR	findCos
ZH, single (LOCAL_ZH)	**SPIRE**	**52.6**	**43.1**	**19.6**	**34.0**	66.2	0.829
	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
	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
	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
	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
	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 8: Per-category retrieval and answer similarity (k=10, paper-level macro-mean, %; findCos = essay-vs-findings BGE-M3 cosine, 0–1). Best eR / workR / secR / MRR per category in **bold**. **SPIRE** leads evidence recall in every category; baselines collapse on CROSS; answer similarity is high and flat (does not track grounding). Naive LLM's retrieval is its from-memory citations resolved to corpus chunks.

Setting	τ_s	τ_c	eR	workR@10	sentR@10	secR@10	MRR@10	nDCG@10
Very permissive	0.70	0.60	57.9	51.2	18.2	23.8	51.0	46.2
Permissive	0.75	0.65	55.7	51.2	17.0	23.5	50.7	45.8
Main	0.80	0.70	53.7	51.2	16.0	23.3	50.5	45.5
Strict	0.85	0.75	51.3	51.2	14.6	22.9	49.6	44.7
Very strict	0.90	0.80	46.5	51.2	12.2	21.8	47.7	42.9

Table 9: Threshold sensitivity of **SPIRE** (`full`) evidence matching (k=10, paper-level macro-mean, %). τ_s is the same-language and τ_c the cross-language BGE-M3 cosine threshold; the main results use $(\tau_s, \tau_c) = (0.80, 0.70)$. Work-level recall is unchanged because it depends on retrieved source works rather than sentence-level cosine matching; sentence recall is the most sensitive metric.

Setting	τ_s	τ_c	Naive LLM	Text RAG	GraphRAG	**SPIRE**
Very permissive	0.70	0.60	18.8	39.2	19.1	**57.9**
Permissive	0.75	0.65	17.9	37.3	17.9	**55.7**
Main	0.80	0.70	16.9	35.2	16.3	**53.7**
Strict	0.85	0.75	15.4	32.7	14.7	**51.3**
Very strict	0.90	0.80	12.6	28.2	11.6	**46.5**

Table 10: Full-pool evidence recall (eR, %) for all systems under the same threshold sweep. The relative ranking is stable across every setting: **SPIRE** leads at all five operating points, and the `Main` row is the midpoint operating point of this diagnostic sweep. Headline retrieval numbers are reported in Table 4. Best per row in **bold**.

weighted Cohen's κ and Gwet's AC2 (Gwet, 2008). They differ only in how they model "chance," and the gap between them is itself informative, so we report both (Table 12, Fig. 3) and treat the prevalence-robust AC2 as primary.

Two notions of chance. κ corrects against the agreement expected if each rater drew scores independently from their own marginal; AC2 corrects against a baseline that does not assume such independence. When scores concentrate at one end of the scale—as here, since every rater scores **SPIRE** high and the weakest baselines low—κ's expected-agreement term inflates and κ is deflated: the well-documented "high agreement, low kappa" paradox (Feinstein and Cicchetti, 1990). Per aspect, AC2 is substantial-to-almost-perfect (LLM–LLM 0.88–0.91, human–human 0.75–0.88, human–LLM 0.66–0.92), while κ on the most top-heavy aspect (human–human Accuracy) drops to 0.35 although the experts differ by more than one point on only 5% of items (within-1 agreement

System	LLM–LLM AC2	LLM–LLM wκ	Human–Human AC2	Human–Human wκ	Human–LLM AC2	Human–LLM wκ
SPIRE	0.95	0.17	0.88	0.14	0.91	0.08
Naive LLM	0.92	0.70	0.79	0.52	0.80	0.53

Table 11: Per-system agreement (item level over the four aspects), on **SPIRE** and Naive LLM (the two systems spanning the score range). AC2 stays high on both, but weighted κ collapses on **SPIRE**—where almost every score is 4–5 and there is little absolute variance to "agree" on beyond chance—and recovers on Naive, whose scores are spread across the scale. This is the prevalence effect that motivates AC2 as the primary statistic.

Aspect	LLM–LLM AC2	LLM–LLM wκ	Human–Human AC2	Human–Human wκ	Human–LLM AC2	Human–LLM wκ
Accuracy	0.88	0.88	0.88	0.35	0.92	0.54
Depth	0.91	0.90	0.77	0.51	0.88	0.82
Coverage	0.90	0.91	0.85	0.78	0.66	0.56
Evidence	0.89	0.88	0.75	0.73	0.86	0.85
Mean	0.90	0.89	0.81	0.59	0.83	0.69

Table 12: Inter-rater agreement per aspect (item level, *free* track). AC2 is Gwet's quadratic-weighted coefficient (our unified statistic); wκ is quadratic-weighted Cohen's κ, shown only to expose the "high agreement, low kappa" paradox under top-heavy marginals (Feinstein and Cicchetti, 1990). LLM–LLM (GPT-5.4 vs. QWEN3.5) pools all four systems (n=400 items/aspect); the human-involving pairs are computed on **SPIRE** and Naive LLM (n=184), the two systems spanning the score range. For human–LLM, each side is the two raters' mean. Both human experts independently rank **SPIRE** first on every aspect.

0.95).

What the agreement is—and is not. Pooled over systems, AC2 is high for every pair (means: LLM–LLM 0.90, human–human 0.81, human–LLM 0.83; Fig. 3a). Much of this is *between-system* agreement: all raters separate **SPIRE** (scored 4–5) from the weakest baselines (scored 1–2), the two diagonal clusters in Fig. 3b. The finer *within-system* absolute agreement is more modest—restricted to a single top-heavy system the score variance nearly vanishes, so κ falls toward chance even though raters seldom differ by more than a point. Table 11 makes this explicit: on **SPIRE**, human–LLM κ=0.08 (AC2 0.91), whereas on Naive LLM, whose scores spread across the scale, κ recovers to 0.53. The judges thus agree strongly on *which* systems are stronger and rarely diverge by more than a point, but their *absolute* per-essay scores agree only moderately; we read the LLM judge as a corroborating proxy for the human ranking, not as a calibrated absolute scorer. Human-involving pairs are computed on **SPIRE** and Naive LLM—the two systems spanning the score range, one top-heavy and one spread across the scale—which isolates the prevalence effect; LLM–LLM uses all four. Both experts and both LLM raters rank **SPIRE** first on all four aspects.

A.6 Component ablations

A.7 Runtime cost

B Robustness to the Generator LLM

To test whether **SPIRE**'s advantage depends on the generator, we repeat the study with GEMINI-3-FLASH in place of DEEPSEEK-V4-FLASH, holding the corpus, substrate, BGE-M3 encoder, retrieval protocol, and rubric judge fixed (§4.2). We report evidence retrieval (Table 15), answer similarity (Table 16), and the comparative LLM judge run head to head against the DEEPSEEK system (Table 17). Across all three measures the ranking of systems is preserved: **SPIRE** leads under the new generator by the same wide margins it shows under the default one.

Retrieval. Under GEMINI-3-FLASH, **SPIRE** keeps a wide retrieval margin over every baseline (Table 15): full-pool evidence recall of 54.1 is more than five times that of flat Text RAG (9.5) and over three times that of GraphRAG (16.4), with the same lead on work, section, and sentence recall and on the ranking metrics. Because retrieval planning is the only stage in which the generator participates, the persistence of this gap shows that **SPIRE**'s evidence-location advantage is a property of the seven-primitive design rather than of a single model.

Answer similarity. Embedding similarity to the reference paper (Table 16) does not separate the strongest systems: **SPIRE**, Naive LLM, and GraphRAG fall within 0.02 on cosine and on BERTScore against both the anchor and the findings text. This reproduces the behaviour observed with DEEPSEEK, where surface similarity saturates and is insensitive to evidence grounding, and it is why retrieval and the rubric judge carry the separation between systems.

LLM judge. **SPIRE** ranks first on every rubric axis under both raters and both generators (Ta-

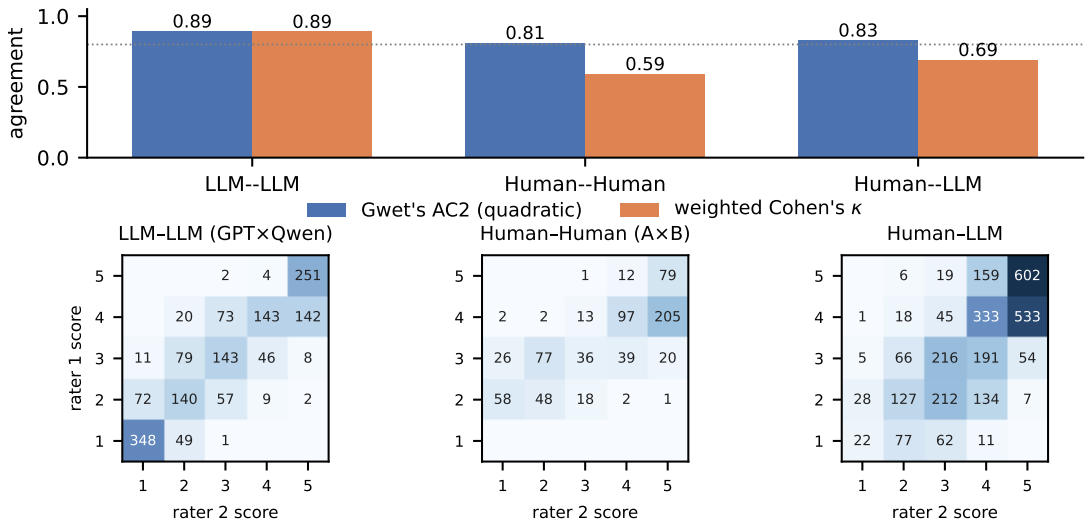

Figure 3: Inter-rater agreement. **(a)** Gwet's AC2 vs. quadratic-weighted Cohen's κ (mean over the four aspects) per rater pair: the two coefficients nearly coincide for LLM–LLM (scores spread across systems) but diverge for the human-involving pairs (top-heavy scores), the kappa paradox. **(b)** Raw 5×5 joint-score counts per rater pair (free track; human-involving pairs on **SPIRE**+Naive LLM). Mass on the diagonal is agreement; the two corner clusters (low for weak baselines, high for **SPIRE**) are the strongly-agreed *ranking* signal, while the off-diagonal spread—widest for human–LLM—is the looser *absolute* agreement.

ble 17), with overall scores of 4.21 and 4.58 under GEMINI. Its lead is widest on EvidenceQuality (4.20 and 4.55, against at most 2.68 and 3.17 for any baseline), the axis that rewards grounded use of primary sources, so the capability the framework is built to deliver is the one that carries over most clearly to the new generator. The two LLM raters agree closely on the free track (Gwet's AC2 of 0.90, in line with the main study; App. A.5). The advantage **SPIRE** shows under DEEPSEEK therefore holds under GEMINI: a consistent first-place ranking whose margin is largest on the evidence-grounding axis at the core of the method.

C Failure Analysis and Qualitative Comparison

Beyond the aggregate and per-category numbers, we examine characteristic *failure modes* of each system on concrete cases, drawing on the human experts' annotations where available (App. H). The contrast is instructive: **SPIRE**'s residual errors are scholarly-*refinement* issues layered on correct primary-source grounding, whereas the baselines fail at grounding itself.

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—accurate, vetted,

Configuration	k	eR	eR@k	workR	sentR	secR	MRR	nDCG	cosSem	R-L	BScore	GPT*	Qwen*
	5	–	19.3	38.5	4.1	10.1	32.7	28.0	–	–	–	–	–
SPIRE (`full`, 7 primitives)	10	44.3	24.0	42.4	5.6	15.3	33.5	28.9	81.8	10.9	84.4	3.44	4.01
	20	–	28.3	45.1	7.8	18.6	33.8	30.3	–	–	–	–	–
w/o one agent (removed from the workflow):													
	5	–	19.6	38.1	3.0	9.6	31.1	28.4	–	–	–	–	–
w/o Annotating ($\mathcal{N}$)	10	41.9	23.3	41.2	4.7	12.7	31.8	29.2	82.2	10.9	84.4	3.16	3.98
	20	–	27.5	44.4	6.4	16.0	32.1	31.2	–	–	–	–	–
	5	–	18.8	38.3	2.4	8.7	32.7	28.5	–	–	–	–	–
w/o Comparing ($\mathcal{P}$)	10	42.0	23.1	41.8	4.1	13.6	33.6	29.4	81.9	10.9	84.4	3.18	3.91
	20	–	27.4	43.3	4.9	17.0	33.8	31.4	–	–	–	–	–
	5	–	18.6	37.4	2.6	8.8	35.6	28.8	–	–	–	–	–
w/o Referring ($\mathcal{R}$)	10	40.9	22.4	40.7	4.3	12.3	36.5	29.7	82.0	11.0	84.5	3.06	3.93
	20	–	26.5	43.3	6.0	15.5	36.8	31.7	–	–	–	–	–
	5	–	16.7	36.7	2.5	8.5	34.0	29.3	–	–	–	–	–
w/o Sampling ($\mathcal{S}$)	10	39.1	20.2	39.8	3.3	11.1	34.9	30.2	82.0	10.8	84.4	3.12	3.92
	20	–	23.9	42.8	4.8	17.1	35.2	32.2	–	–	–	–	–
	5	–	12.1	29.8	1.9	7.1	21.4	18.3	–	–	–	–	–
w/o Illustrating ($\mathcal{I}$)	10	27.8	14.6	32.4	2.9	10.0	21.9	18.9	82.0	10.9	84.4	3.07	3.74
	20	–	17.3	34.4	3.9	12.6	22.1	20.2	–	–	–	–	–
Retrieval tier disabled:													
	5	–	14.0	28.3	1.8	7.9	26.4	22.2	–	–	–	–	–
w/o cross-context cluster	10	32.5	16.9	31.7	3.6	11.5	27.1	22.9	70.9	8.1	83.0	3.13	3.77
	20	–	20.8	34.4	4.7	13.7	27.3	24.4	–	–	–	–	–
	5	–	10.7	25.7	1.7	7.0	23.5	19.9	–	–	–	–	–
w/o intra-context community	10	33.5	12.8	28.0	3.1	9.9	24.1	20.5	63.4	5.7	82.2	3.10	3.51
	20	–	16.4	31.1	4.2	12.2	24.3	21.9	–	–	–	–	–
	5	–	7.5	20.6	1.7	6.0	22.0	18.6	–	–	–	–	–
w/o both (local only)	10	28.5	9.4	22.6	2.4	8.1	22.5	19.2	63.3	5.9	82.2	3.05	3.20
	20	–	11.3	25.5	3.0	9.6	22.7	20.5	–	–	–	–	–

Table 13: Component-ablation sweep, $k \in \{5, 10, 20\}$ (paper-level macro-mean, %). The main paper reports the k=10 slice in Table 6; this appendix table reports all three evaluated k values for completeness. eR (full-pool, rank-independent), the answer-similarity columns (cosSem / R-L / BScore vs. the paper findings), and the two LLM-judge overall-score columns (GPT-5.4, Qwen3.5; mean of the four aspects, 1–5) carry no k and appear on the k=10 row only. nDCG/MRR are computed on sectionR. All metrics are measured on the full 406-paper benchmark (eR is k-invariant).

System / run	lat. (min)	tokens	steps
Naive LLM†	0.22	≈1k	1
Text RAG	1.39	≈36k	2
GraphRAG	1.06	≈37k	2
SPIRE (`full`)	4.37	505k	22

Table 14: Runtime cost, per-paper means over successful runs, same generator (DEEPSEEK-V4-FLASH); tokens are total prompt+completion. Baseline token counts are estimated from the retrieved context plus answer length (≈; they were not logged), and "steps" counts LLM calls: Naive answers in a single call; Text RAG and GraphRAG add one query-planning call before synthesis. For **SPIRE**, "agent steps" counts agent invocations in the trace (a proxy for LLM calls; each may issue more than one): it runs all seven agents with up to two reflection rounds (≈22 steps). **SPIRE**'s provenance gains (Table 4) come at a deliberate cost: it spends ∼3× the latency of Text RAG and an order of magnitude more tokens, the price of agentic close reading. †Naive LLM uses no retrieval.

System	eR	eR@10	workR@10	sentR@10	secR@10	MRR@10	nDCG@10
DEEPSEEK-V4-FLASH (default, reproduces Table 4)							
Naive LLM	14.3	14.3	17.4	3.6	4.4	15.7	15.0
Text RAG	22.4	14.5	12.5	2.7	3.5	6.5	4.0
GraphRAG	12.4	12.5	13.2	2.3	3.6	6.2	8.9
SPIRE (`full`)	**44.3**	**24.0**	**42.4**	**5.6**	**15.3**	**33.5**	**28.9**
GEMINI-3-FLASH (robustness re-run)							
Naive LLM	14.0	14.3	16.8	3.5	4.3	15.2	14.4
Text RAG (flat)	9.5	4.2	0.4	0.2	0.1	0.4	0.1
GraphRAG	16.4	16.4	12.8	2.5	6.1	7.6	7.4
SPIRE (`full`)	**54.1**	**34.9**	**51.2**	**8.4**	**23.1**	**52.7**	**53.0**

Table 15: Evidence retrieval under the two generators (paper-level macro-mean, %); best per column within each generator block in **bold**. eR is rank-independent full-pool recall.

System	Anchor			Findings		
	cos	R-L	BERT	cos	R-L	BERT
SPIRE	0.778	0.102	0.841	0.814	0.102	0.843
Naive LLM	0.786	0.124	0.851	0.800	0.077	0.851
Text RAG	0.642	0.064	0.830	0.636	0.032	0.827
GraphRAG	0.775	0.109	0.837	0.794	0.088	0.839

Table 16: Answer similarity to the reference paper under GEMINI-3-FLASH: cosine, ROUGE-L (R-L), and BERTScore (BERT) against the anchor sentence and the full findings text.

and citable—whereas an ad hoc model translation is none of these. 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.

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*)—nothing 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—the very failure that motivates evidence-grounded retrieval.

In sum, **SPIRE**'s residual errors presuppose it found and quoted the right primary sources (it then mis-periodises a term, or self-translates), whereas the baselines fail upstream—retrieving wrong or secondary material, or nothing usable. This is why the gap is largest precisely on EvidenceQuality (Table 5) and on work/section recall (Table 4).

D Graph, Community and Cluster Substrate

D.1 Graph-extraction quality audit

Because the multi-scale substrate (§3.3) is built on LLM-extracted entities, relations, and supporting evidence, we run an intrinsic expert audit of extraction quality, complementing the end-to-end retrieval and answer metrics of §5. We randomly sample 5% of chunks, balanced across Chinese and Latin, and ask two domain experts (one in classical Chinese, one in Latin philology) to rate each extraction on seven dimensions: Entity Comprehensiveness, Entity Accuracy, Entity Usability, Relation Comprehensiveness, Relation Accuracy, Relation Usability, and Evidence Accuracy. Because intellectual-graph extraction from classical texts has no single exhaustive gold graph (experts may legitimately disagree about which concepts and relations are interpretively salient), we use holistic expert ratings on a 5-point Likert scale rather than precision/recall against a fixed label set.

Table 18 reports model-level means under the identical extraction prompt and JSON schema (App. G). GPT-5.4 gives the strongest proprietary upper bound, while DEEPSEEK-V4-FLASH (thinking) reaches near-frontier quality and is strongest or tied on Relation Comprehensiveness, Entity Usability, and Evidence Accuracy. Relative to DeepSeek-V3 and DeepSeek-R1 it substantially improves entity accuracy, relation usability, and format stability. Compared with GPT-5.4 it offers a better cost–availability trade-off while preserving the extrac-

System	Rater A (GPT-5.4) Acc	Dep	Cov	EQ	Ov	Rater B (Qwen3.5) Acc	Dep	Cov	EQ	Ov
DEEPSEEK-V4-FLASH										
SPIRE	**4.49**	**4.44**	**4.75**	**4.33**	**4.50**	**4.88**	**4.86**	**4.87**	**4.79**	**4.85**
Naive LLM	3.58	2.50	3.29	2.10	2.87	3.87	2.10	3.01	1.90	2.72
Text RAG	4.08	3.69	4.21	3.51	3.87	4.29	3.45	4.22	3.20	3.79
GraphRAG	3.43	3.07	3.76	2.68	3.23	3.77	2.92	3.71	2.32	3.18
GEMINI-3-FLASH										
SPIRE	**4.25**	**4.03**	**4.34**	**4.20**	**4.21**	**4.59**	**4.56**	**4.60**	**4.55**	**4.58**
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 17: Comparative LLM-judge scores on the free track (humanities rubric, 1 to 5; Acc: AnswerAccuracy, Dep: ArgumentDepth, Cov: CoverageCompleteness, EQ: EvidenceQuality, Ov: overall mean) on the shared human-rated question set (n=100); best per column within each generator block in **bold**.

tion quality needed for downstream retrieval, and it stays in the same model family as the generator used elsewhere in the pipeline (§4.2). We therefore use DEEPSEEK-V4-FLASH (thinking) for production graph construction.

D.2 Graph and cluster artefacts per scale

Table 19 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, so cross-context recurrence should not be equated with automatic cross-language alignment. Cross-tradition alignment is therefore performed at the agent level by the Comparing primitive rather than expected from the vector index.

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)

```
paper: "Pietas in pro Sexto Roscio of Cicero and Confucian xiao"
paper_language: en   literature_languages: zh_classical + la_classical

research_question:
  How does Cicero's pietas in pro Sexto Roscio relate to Confucian xiao,
  and what does comparison show about filial duty and social order?

research_findings (excerpt):
  Parricide is treated as the ultimate filial impiety; Cicero rejects the
  charge by arguing ex persona, ex causa, and ex facto ipso.

evidences (2 of 13 shown):
  E01  Pro Roscio Amerino  loc 37  translated_quote  central
       "Sextus Roscius stands accused of the murder of his father. ..."
  E04  Pro Roscio Amerino  loc 69  original_quote  supporting
       "quod in impios singulare supplicium invenerunt"
```

Extracted graph. Per-chunk DEEPSEEK-V4-FLASH output (§3.3), abridged to a few entities/relations; Chinese glyphs are romanized to match the paper's convention.

Extracted graph – Latin chunk (Ammianus Marcellinus, Res Gestae)

```
{"entities":[
  {"type":"Person","title_modern":"Nohodares",
   "description_modern":"A Persian nobleman who planned an invasion of Mesopotamia.",
   "evidence_spans":[2]},
  {"type":"Concept","domain":"Politics & Governance","title_modern":"Saracen customs",
   "description_original":"Saracenorum irruptiones et mores",
   "description_modern":"The nomadic, predatory life and military habits of the Saracens.",
   "evidence_spans":[9,10,11,12,13,14,15]},
  {"type":"Concept","title_modern":"Persian stratagem",
   "description_original":"Persarum commentum irritum",
   "description_modern":"A failed Persian military plan or ruse.","evidence_spans":[1]}],
 "relations":[
  {"source":"Ammianus Marcellinus","target":"Saracen customs","relation_name":"describes",
   "description":"Ammianus gives an ethnographic account of Saracen customs.",
   "evidence_spans":[9,...,15]},
  {"source":"Nohodares","target":"Persian stratagem","relation_name":"devised",
   "description":"Nohodares planned the invasion of Batnae during its annual fair.",
   "evidence_spans":[2,...,7]}]}
```

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

```
{"entities":[
  {"type":"Concept","domain":"Religion & Spirituality","title_modern":"fangbian (skillful means)",
   "description_modern":"A Buddhist term for expedient teaching; here the expedient by which
    earlier worthies reconciled the Three Teachings.","evidence_spans":[2,19,20,28]},
  {"type":"Concept","title_modern":"sanjiao (Three Teachings)",
   "description_modern":"Confucianism, Buddhism, and Daoism; the text argues they can be unified.",
   "evidence_spans":[25,34]},
  {"type":"Person","title_modern":"Cheng Yi",
   "description_modern":"Northern-Song Neo-Confucian; here decries Buddhism as heterodox and monstrous.",
   "evidence_spans":[59,90]}],
 "relations":[
  {"source":"Liu Mi","target":"sanjiao","relation_name":"argues_for_unity",
   "description":"Liu Mi holds the Three Teachings can be merged into one.","evidence_spans":[25,26,27]},
  {"source":"Liu Mi","target":"Cheng Yi","relation_name":"refutes",
```

Model	Ent.Comp	Ent.Acc	Ent.Use	Rel.Comp	Rel.Acc	Rel.Use	Ev.Acc
GPT-5.4	**4.86**	**4.84**	4.82	4.82	**4.86**	**4.88**	4.84
DeepSeek-V4-Flash (thinking)	4.72	4.76	**4.84**	**4.88**	4.82	4.84	**4.86**
DeepSeek-V3	4.00	4.16	4.60	4.84	4.48	3.64	4.52
DeepSeek-R1	3.86	4.08	4.36	3.88	3.74	3.32	4.86
Qwen3-32B	4.04	4.80	4.84	4.22	4.80	**4.88**	4.00

Table 18: Expert validation of LLM-based graph extraction on a 5-point Likert scale, means over sampled Chinese and Latin chunks under one shared extraction prompt and schema (App. G). Dimensions: Entity / Relation Comprehensiveness (Comp), Accuracy (Acc), Usability (Use), and Evidence Accuracy (Ev.Acc). Best per column in **bold**. DEEPSEEK-V4-FLASH (thinking), used in production, is at or near the proprietary GPT-5.4 upper bound.

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 19: 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).

```
    "description":"Liu Mi rebuts Cheng Yi, noting the Confucian classics also contain strange tales.",
    "evidence_spans":[66,...,92]}]}
```

F Agent Prompts and Deterministic Rules

Every agent's system prompt is given verbatim below (the longest, most repetitive worked-example enumerations are elided as `[...]` and released in full with the code). The ablations (§5.3) *remove* an agent from the workflow.

Discovering – query-parse system prompt

```
You are the PARSE step of the Discovering primitive in SPIRE, an AI multi-agent system for traditional humanities research (literature, history, philosophy), focused on Ancient Chinese and Latin/Greco-Roman studies. You receive a research question from a humanities paper. Analyse it, then infer parameters a downstream orchestrator uses to compose and run the research agents.

OUTPUT: a single JSON object.
{
  "sub_questions": [
    {"text": "<one sub-question, in English>",
     "granularity": ["chunk|fine|mid|broad|tradition"],
     "perspective": ["local|explicit|implicit"],
     "domain_prior": ["<research-field labels>"]}
  ],
  "languages": ["zh|la"],
  "comparison": ["none|intra_text|intra_author|intra_tradition|cross_tradition"],
  "entity_anchors": ["<persons/concepts EXPLICITLY in the question>"],
  "work_anchors": ["<classical texts EXPLICITLY in the question>"],
  "retrieval_expansion_hints": {
    "related_concepts": [...], "candidate_primary_works": [...],
    "historical_periods": [...], "synonyms_or_translations": [...],
    "native_terms": {"zh": ["<6-12 Hanzi keywords>"], "la": ["<6-12 Latin keywords>"]}
  },
  "analytical_lens": null,
  "answer_language": "en|zh|la",
  "reasoning": "<one sentence>"
}

sub_questions: decompose aggressively for retrieval; each sub-question is a separate embedding
query. Default 2-4 unless genuinely single-pointed; decompose on any conjunction, comparison
signal, transformation chain, enumeration of aspects, relation between concepts, or multiple
wh-words. Cap 4. [worked DO/DON'T examples ...]
granularity = level of abstraction (chunk|fine|mid|broad|tradition).
perspective = local (the passages themselves) | explicit (relations
the texts attest) | implicit (concepts kindred in meaning with no attested link). domain_prior =
soft field hint, not a hard filter.
languages = source-corpus language(s), not the answer language.
comparison: cross_tradition for zh+la; intra_tradition for one tradition across periods/authors;
```

```
intra_text/intra_author within one book/author.
entity_anchors/work_anchors: LITERAL only.
retrieval_expansion_hints: soft, retrieval-only,
never cited.
answer_language = the language the question is
asked in. [trigger-word
notes ...]
```

Annotating – system prompt

```
You are a humanities research assistant performing
 scholarly annotation. Given a research
question (decomposed into sub-questions) and
retrieved passages, annotate each passage:
1. INTERPRETATION: what the passage tells us in
relation to the
   question (1-2 sentences of interpretive
significance, not a summary).
2. SUB_QUESTION: which sub-question id (q1, q2,
...) it best bears on;
   "general" if several or none.
3. CONNECTIONS: references it makes to other works
/authors/concepts
   that also appear elsewhere in the batch.
4. RELEVANCE_RANK: 1-5 (5 = directly addresses the
 core with specific
   textual evidence; 1 = tangential background).
Output VALID JSON only: a single object with an "
annotations" array of
{text_unit_id, interpretation, sub_question,
connections[],
relevance_rank}. Do NOT return a bare array.
```

Comparing – system prompt

```
You are performing a scholarly comparison of
evidence. Research question: {query}; Comparison
type: {comparison_kind}. Evidence side A ({
side_a_label}); evidence side B ({side_b_label}).
Perform a structured comparison (concept-cluster
comparative method): identify parallel concepts
across the sides, structural divergences,
unilateral concepts (present on one side, absent
on
the other), and inverted emphases on a shared
theme. Treat any apparent analogue as a scholarly
judgement, not an established equivalence.
Output VALID JSON only: {parallel_concepts[],
structural_divergences[],
unilateral_concepts[], inverted_emphases[],
comparison_summary}.

[cross-tradition suffix] SPIRE does NOT pre-
compute cross-tradition
links; alignment is your job from the side-by-side
 pools. Map each side's concept cluster from
its own evidence pool; give candidate analogues (e.
g. ren <-> humanitas), divergences,
unilateral concepts, inverted emphases. BALANCE (
hard rule): every side with Source Text
Passages must contribute >=1 verbatim, [E<id>]-
tagged quotation; a
one-sided essay is a FAILED synthesis; state any
shortfall under limitations/source_coverage.
```

Reflection – prompt

```
You are a humanities scholar reviewing a draft
against the question. Assess: (1) comprehensive
coverage of the question and sub-questions; (2)
obvious gaps; (3) evidence balance across
traditions/sources; (4) unsupported claims.
Output VALID JSON only: {comprehensive, gaps[{
description,
suggested_action}], balance_issue,
unsupported_claims[], verdict
("pass"|"re-discover"|"minor-gaps"), summary}. A
re-discover query must
be a concrete corpus search (work/author/concept/
native term); never a
workflow instruction such as "regenerate" or "find
 more evidence".
```

Referring – system prompt

```
You are validating the provenance and connections
of scholarly evidence.
For each passage assess: (1) PROVENANCE
completeness (author, work,
chapter); (2) CONNECTIONS it explicitly makes (
citation/allusion/ response/commentary); (3)
SECONDARY RETRIEVAL worth running; (4) OUT-OF-
CORPUS WORKS: works named in the question or
passages that no retrieved passage represents.
Inputs: research question; works named in the
question; distinct works in the retrieved evidence
; the evidence passages.
Output VALID JSON only: {provenance_checks[],
discovered_connections[],
secondary_retrieval_suggestions[{query,reason,
language}],
out_of_corpus_works[]}.
```

Sampling – system prompt

```
You are selecting and profiling evidence for a
scholarly research output. Tasks: (1) SELECT 8-15
passages that together cover the sub-questions;
(2) DISTRIBUTION by tradition (zh/la),
sub-question, source diversity; (3) GAPS (a sub-
question with little support, a tradition
under-represented for a cross-tradition question,
over-concentration in one work).
Output VALID JSON only: {selected_ids[],
distribution{by_tradition,
by_sub_question, unique_works, work_list},
coverage_gaps[{description, severity}],
selection_rationale}.
```

Illustrating+Representing – synthesis system prompt

```
You are the synthesis stage of SPIRE, a research
assistant for close reading of Chinese and
Latin/Greco-Roman classical texts in intellectual
history.

Task: given a research question, the retrieved
classical-text passages,
and your scholarly knowledge, write a humanities
research essay. SPIRE calls you in two rounds:
round 1 write the essay as prose; round 2 extract
structured findings and evidences from the
finished essay.

Writing: write as a scholar would, in the question
's language; identify the object of inquiry,
articulate the question, situate it in the
retrieved primary sources and recognised secondary
scholarship, develop interpretive analysis with
claims supported by primary-source quotation and
argument, and reach a synthesis. Treat the
retrieved Source Text Passages as the principal
evidence; scholarly knowledge may supplement where
 retrieval is partial, but may NOT replace an
entire side of a comparison with from-memory
paraphrase; if a side's passages are too thin, say
so as a limitation.

Citation: inline format for classical texts is
  "original quotation" (Work, Author, Section) [E<
id>]
Every verbatim classical quotation MUST be
followed by an [E<id>] tag;
an untagged quotation is invalid. Every structured
 evidences[] entry
```

```
(with a text_unit_id) MUST come from a retrieved Source Text Passage; do not fabricate
original-language wording or locators from memory. Auxiliary references (further loci classici
outside the corpus,
secondary scholarship) stay in the prose voice, not in evidences[].
Never write a text_unit_id or an entity/community/cluster title into the prose as if it were a
document.

Two rounds: round 1 returns the essay only (prose with inline
[E<id>]-tagged quotation); round 2 returns the audit JSON only
(findings[], evidences[], source_coverage, suggested_followup_queries),
reusing the same E<id> values. Evidence forms supplied: (a) Source Text Passages (the only
source of hard text_unit_id citations); (b) community summaries; (c) semantic-family (cluster)
summaries; (d) entity/relationship context. Use (b)-(d) as scaffolding only.
```

G Graph-Extraction and Report Prompts

The entity-relation graph (§3.3) is extracted per chunk, and every Leiden community and HDBSCAN semantic-family cluster is summarized into a report, by DEEPSEEK-V4-FLASH (thinking; temperature 0.1). The system prompts are below, lightly condensed; Chinese example glyphs are romanized to match the paper's convention, and the literal prompts (with original-language text) ship with the code. A user may replace the extraction schema for another field.

Graph-extraction system prompt

```
You are a rigorous classical humanities scholar extracting a culture knowledge graph from
Chinese or Latin classical texts. Extract only the most important entities and relations useful
for intellectual-history and cross-cultural analysis. Quality matters more than coverage; when
in doubt, do NOT extract.

=== ENTITY TYPES (exactly three) ===
1. Concept -- a recurring intellectual touchstone in a tradition
   (concept, doctrine, motif, value, or problem recurring across texts).
   Chinese e.g. ren, yi, li, dao, tian, ming, xing-shan, xing-e,
   si-duan, gewu-zhizhi, wu-wei, zhongyong, kong, yuanqi, fo-xing.
   Latin/Greco-Roman e.g. iustitia, virtus, fides, pietas, officium,
   prudentia, ratio, logos, natura, libertas, res publica, lex naturae.
   NOT: idioms, proverbs, narrative phrases, concrete things, places,
   institutions, sentence fragments, or loosely used ordinary words.
2. Person -- a named individual who proposes, defines, debates,
   transmits, criticizes, or exemplifies a Concept (Confucius, Mencius,
   Zhu Xi, Cicero, Seneca, Augustine, Aristotle).
3. Work -- a named text/treatise/poem/collection that carries, defines,
   records, or debates a Concept (Lunyu, Mengzi, Daxue, De Officiis).

=== ENTITY FIELDS (all required) === type (Concept|Person|Work); domain (Ethics & Morality |
Politics & Governance | Religion & Spirituality | Metaphysics & Epistemology | Natural
Philosophy & Science | Literature & Aesthetics | Other; Person and Work always Other);
title_original (source-language form, no
brackets/quotes, <=10 words); title_modern (standard academic name;
Simplified Chinese for zh, English for la); description_original (short phrase from / closely
paraphrasing the text); description_modern (1-2 sentence faithful explanation); evidence_spans
(sentence indices only, e.g. [0,2]; never a text_unit_id).

=== RELATIONS (all required) === source, target (title_modern of extracted entities),
source_type, target_type, relation_name (concise English verb phrase), description (1-2
sentences; zh for Chinese chunks, en for Latin), evidence_spans.
Good relation_name: authors, records, defines_as, argues_for,
criticizes, exemplifies, is_foundation_of, is_aspect_of, is_opposite_of, develops_into,
transmits, compares_with, is_condition_for, is_core_of. Bad: mentions, discusses, relates_to.
source and target MUST appear in the entities array.

=== LIMITS & OUTPUT ===
<=30 entities and <=50 relations per chunk. Return ONLY a JSON object
with keys "entities" and "relations" (both arrays);
{"entities": [], "relations": []} if none. No markdown, no
chain-of-thought. The user message supplies TEXT UNIT ID, LANGUAGE, TEXT CONTENT, and
SENTENCE_SPLITS (indexed JSON).
```

Graph-extraction few-shot examples (appended to the prompt)

```
### Chinese example (Analects 1.2; sentences romanized)
Sentences: {
  0: "Youzi said: one who is filial and fraternal yet defies superiors
      is rare.",
  1: "The junzi works at the root; the root established, the Way grows.",
  2: "Filial-fraternal conduct (xiao-ti) is the root of ren."
}
Output:
{"entities":[
  {"type":"Person","domain":"Other","title_original":"Youzi",
   "title_modern":"Youzi","description_original":"Youzi said",
   "description_modern":"A disciple of Confucius; here relates xiao-ti
   to ren.","evidence_spans":[0,2]},
  {"type":"Concept","domain":"Ethics & Morality",
   "title_original":"xiao-ti","title_modern":"xiao-ti",
   "description_original":"xiao-ti is the root","description_modern":
   "Filial piety and fraternal respect as an ethical virtue.",
   "evidence_spans":[0,2]},
  {"type":"Concept","domain":"Ethics & Morality","title_original":"ren",
   "title_modern":"ren","description_original":"the root of ren",
   "description_modern":"The core Confucian ethical concept.",
   "evidence_spans":[2]}
],"relations":[
  {"source":"Youzi","target":"xiao-ti","source_type":"Person",
   "target_type":"Concept","relation_name":"argues_for",
```

```
    "description":"Youzi holds xiao-ti the ethical root of the junzi.",
    "evidence_spans":[0,2]},
  {"source":"xiao-ti","target":"ren","source_type":"Concept",
    "target_type":"Concept","relation_name":"is_foundation_of",
    "description":"The text states xiao-ti is the root of ren.",
    "evidence_spans":[2]}
]}

### Latin example (Cicero, De Officiis)
Sentences: {
  0: "Marcus Tullius Cicero librum De Officiis ad filium scripsit.",
  1: "Fundamentum autem est iustitiae fides, id est dictorum
      conventorumque constantia et veritas."
}
Output:
{"entities":[
  {"type":"Person","domain":"Other",
    "title_original":"Marcus Tullius Cicero","title_modern":"Cicero",
    "description_original":"librum De Officiis ad filium scripsit",
    "description_modern":"Roman orator, statesman, and philosopher of
    the late Republic.","evidence_spans":[0]},
  {"type":"Concept","domain":"Ethics & Morality",
    "title_original":"fides","title_modern":"Trustworthiness",
    "description_original":"dictorum conventorumque constantia et
    veritas","description_modern":"The virtue of keeping one's word and
    honoring agreements.","evidence_spans":[1]},
  {"type":"Concept","domain":"Ethics & Morality",
    "title_original":"iustitia","title_modern":"Justice",
    "description_original":"Fundamentum autem est iustitiae fides",
    "description_modern":"A moral virtue whose foundation is
    trustworthiness.","evidence_spans":[1]}
],"relations":[
  {"source":"Trustworthiness","target":"Justice",
    "source_type":"Concept","target_type":"Concept",
    "relation_name":"is_foundation_of","description":"Cicero states
    trustworthiness is the foundation of justice.","evidence_spans":[1]}
]}
```

Community-report system prompt (intra-context tier)

```
You are a classical humanities scholar producing a substantive scholarly report on an
intellectual community detected from a corpus of Chinese or Latin classical texts. You receive
(a) the community's member entities (Concept, Person, Work) with descriptions and (b) the
relationships among them. The report is used for SPIRE retrieval, cross-tradition comparison,
and citation -- produce something rich, not a one-paragraph summary.

OUTPUT: a JSON object, all fields required:
- title: canonical 3-10 word title of what UNIFIES the community
  (Chinese for zh communities, English for la).
- summary: 3-6 sentences -- the binding theme, the kind of intellectual
  move (school, doctrine, controversy, ritual cluster, technical
  vocabulary, period), and the dominant domain.
- narrative: 3-6 paragraphs -- (1) core thesis / why it cohered;
  (2) 5-8 key entities and their roles (highest-degree members);
  (3) relationship structure (critique chains, commentary lineages,
  is_part_of hierarchies); (4) historical / intellectual context;
  (5, opt., size>30) sub-clusters; (6, opt.) reception / significance.
  Small communities (<10): 1-2 paragraphs.
- key_findings: 5-10 {name (<=8 words), description (1-3 sentences
  citing concrete entities)}, ordered by importance.
- top_entities: 5-10 {id (verbatim from input), title_modern, type,
  role (central|supporting|peripheral), role_description}.
- themes: 5-10 short keyword retrieval tags (<=6 words; zh for zh,
  en for la; no duplicate of the title).
- rating: integer 1-10 (10 = central pillar of the tradition;
  7-8 = important; 4-5 = specialized; 1-3 = noisy co-occurrence).
- rating_explanation: 1-2 sentences vs. other communities in the
  same tradition.
Stay grounded in the given entities/relations (invent nothing); use canonical academic
terminology; surface the highest-degree members in top_entities using their id verbatim; scale
narrative to size (~150-300 / ~400-700 / ~700-1200 words); output ONLY the JSON.
```

Semantic-family (cluster) report system prompt (cross-context tier)

```
You are a humanities scholar producing a SEMANTIC-FAMILY report on a cluster of concepts formed
by EMBEDDING-SPACE PROXIMITY of entity descriptions, NOT by shared text-network membership --
members may come from different works, authors, and periods without ever being mentioned
together. Identify the latent conceptual affinity (a Lovejovian
"unit-idea") that pulled them together.

OUTPUT_LANGUAGE (given in the user prompt) is strict: zh -> all free-text fields in Simplified
Chinese (entity names as given); la -> clean academic English (romanize / gloss non-English
names on first mention). Do not mix languages.

Interpretive stance (Lovejoy "history of ideas", within a tradition):
(A) semantic affinity != textual connection -- do not assume members cite each other or the same
authors; (B) name the shared conceptual core (e.g. "innate moral capacity", "ritual order",
"natural law");
(C) note variants / developments of the unit-idea across authors and periods (e.g. for a Chinese
cluster: xing-shan in Mencius -> liang-zhi in Wang Yangming); (D) contrast with graph
communities -- state what this cluster surfaces that a citation / co-occurrence network misses.

OUTPUT: the SAME JSON schema as the community report (title; summary; narrative of 3-6
paragraphs: core thesis / key entities & variants / structure within the family / what this
reveals vs. a graph community / opt. sub-clusters / opt. reception; key_findings; top_entities;
themes; rating 1-10; rating_explanation), all free text in OUTPUT_LANGUAGE. Stay grounded; use
entity ids verbatim; scale length to cluster size; output ONLY the JSON, single-language per
OUTPUT_LANGUAGE.
```

H 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 four 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)

```
You are a humanities scholar reviewing research-assistance material (not a publication-ready
paper). You are given the research question, the paper title that produced it (background only,
NOT a thesis to replicate), and the essay under review. In the comparative protocol you see
SEVERAL anonymised candidate essays for the SAME question in random order; score EACH
independently on the same 1-5 rubric. Length is not rewarded.

1. AnswerAccuracy -- does the essay precisely address the relation the
   question asks (how / why / in what sense / through what mechanism)
   with textually sound claims, not merely list related concepts?
   1 = off-topic / contradicts the texts; 2 = misreads the core
   concept; 3 = relevant but generic or half-answers the key relation
   (topical adequacy CAPS HERE); 4 = accurate, minor simplification;
   5 = precise on the core relation, comparanda kept distinct.

2. ArgumentDepth -- is the argument driven by specific textual detail
   and developed in layers, not paraphrastic summary?
   1 = no real argument; 2 = assertion/paraphrase only; 3 = a thesis
   but mainly generalisation or restatement (SUMMARY CAPS HERE);
   4 = structured multi-step argument, several moves text-driven;
   5 = layered analysis deriving claims from close reading of specific
   passages.

3. CoverageCompleteness -- does it cover the sub-tasks the question
   entails? (A comparison requires both sides + similarity + difference
   + the limits of the comparison.)
   1 = misses major dimensions; 2 = one side only; 3 = main question
   answered but a key sub-dimension thin (a single general point CAPS
   HERE); 4 = main dimensions covered, minor omissions; 5 = thorough
   and handles nuance / limitations.

4. EvidenceQuality -- how well is the answer grounded in specific
   primary-text evidence: density of direct quotation, specificity of
   citation, close reading of the source material, versus vague
   work-name dropping or hand-waving paraphrase? (Faithfulness floor:
   treat a citation as fabricated only if the quoted words clearly do
   not belong to the cited work or are mis-attributed; this is rare and
   not the main concern.)
   1 = essentially no primary-text evidence or clear fabrication;
   2 = evidence very sparse, loosely connected; 3 = some real evidence
   but sparse/generic or bare work-name mention (~no direct quotation
   CAPS HERE); 4 = key claims backed by specific, accurate quotation;
   5 = quotation specific and tightly integrated with close-reading
   analysis throughout.

Output VALID JSON only. Single essay:
{"AnswerAccuracy": <1-5>, "ArgumentDepth": <1-5>,
 "CoverageCompleteness": <1-5>, "EvidenceQuality": <1-5>,
 "rationale": "<one short sentence per axis, in order>"}
Comparative (N essays):
{"scores": {"Answer 1": {"AnswerAccuracy": <1-5>, ...},
 "Answer 2": {...}, ...},
 "comparative_note": "<which answer is most grounded in direct
 primary-text evidence, which is mostly generic summary>"}
```

I Fragment-Level Qualitative Comparison Figures

Reading guide. Green marks direct primary evidence or precise textual citation; cyan marks evidence-to-claim reasoning; blue marks broad thesis statements; orange marks source mentions or retrieval locators; purple marks graph/community-level summaries; red marks limitations or retrieval imbalance. Ellipses indicate omitted surrounding text from the original system outputs. The fragments below show representative excerpts from the complete **SPIRE** outputs used in the qualitative analysis.

Figure A. 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？

Caption. Fragment-level comparison on a Chinese intellectual-history question. **SPIRE** differs from the baselines by using primary-text evidence not merely as support, but as the basis for finer-grained conceptual distinctions between Cheng Hao and Cheng Yi.

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 textbook contrast, Text RAG adds relevant quotations, and GraphRAG broadens the conceptual map. **SPIRE** instead 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

程颢的修养工夫以“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 contrast is not merely whether a system cites sources, but what it does with them. **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.

Figure B. 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?

Caption. Fragment-level comparison on a Latin close-reading question. **SPIRE** reconstructs Cicero's argument by linking Latin evidence to a precise claim about correlation, causation, and explanatory adequacy.

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 eventura"* (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, showing that the correlation/causation distinction 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 can discuss evidence and explanation, but 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. That balanced reconstruction makes the later critique sharper because Marcus is shown to challenge a plausible evidential practice rather than a caricature.