TRACE: Accountable Agentic Retrieval for Source Discovery in Digital Archives

Donghan Bian, Marie Puren, Florian Cafiero

TRACE: Accountable Agentic Retrieval for Source Discovery in Digital Archives

Donghan Bian
École nationale des chartes – PSL, Centre Jean-Mabillon
Paris, France
EPITA, LRE
Le Kremlin-Bicêtre, France

Marie Puren
EPITA, LRE
Le Kremlin-Bicêtre, France
École nationale des chartes – PSL, Centre Jean-Mabillon
Paris, France

Florian Cafiero
EPITA, LRE
Le Kremlin-Bicêtre, France
CDHM, Geneva Graduate Institute
Geneva, Switzerland

Abstract

Historical archives pose a difficult retrieval problem for retrieval-augmented generation systems: documents are OCR-degraded, heterogeneous across genres and sources, and require strong source traceability for scholarly and institutional use. We introduce TRACE, a training-free agentic retrieval framework designed for accountable source discovery over historical corpora. The system was developed in the context of DECIDON, an interdisciplinary project on the circulation of political discourse between parliamentary debates and the press during the French Third Republic, involving digitised historical collections and institutional use cases. The prototype is currently deployed internally within the project and accessible to 24 researchers across six partner institutions. We evaluate TRACE on HistoriQA-ThirdRepublic, a benchmark of 1,752 French historical questions over parliamentary debates and newspapers from 1887, with documents derived from Bibliothèque nationale de France digitised collections. TRACE achieves R@10 = 0.856 and MRR = 0.653, outperforming sparse, dense, graph-based, and agentic RAG baselines, with the largest gains on multi-hop and cross-corpus questions. At approximately $0.02 per question under the default hosted inference configuration, TRACE also remains economically feasible for heritage institutions, laboratories or companies that cannot rely on costly local GPU infrastructure. These results suggest that, for large digital libraries and archives, retrieval accountability and corpus-aware agent design can provide a practical alternative to heavier training-based or graph-construction approaches.

CCS Concepts

• **Information systems** → **Question answering**; **Information retrieval**; • **Computing methodologies** → *Natural language processing*.

Keywords

agentic retrieval, retrieval-augmented generation, historical corpora, digital archives, source discovery, OCR-degraded text, multi-hop retrieval

ACM Reference Format:
Donghan Bian, Marie Puren, and Florian Cafiero. 2026. TRACE: Accountable Agentic Retrieval for Source Discovery in Digital Archives. In *Proceedings of 35th ACM International Conference on Information and Knowledge Management (CIKM '26).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Archival documents have long served as a cornerstone of historical research [6]. With the convergence of computer science and digital humanities, large volumes of paper-based materials have been digitised, making large-scale search increasingly feasible. Yet deploying retrieval-augmented generation (RAG) [20] in historical research settings remains hindered by three compounding challenges. At the *technical* level, high OCR error rates introduce semantic drift during embedding and degrade retrieval precision [7, 36]. At the *query* level, historical questions routinely span multiple heterogeneous sources, extended temporal scopes, or require cross-document synthesis rather than single-fact lookup. At the *methodological* level, most RAG systems optimise for generation quality [25, 33], whereas historical scholarship imposes strict source traceability, interpretive transparency, and the possibility of human intervention at every retrieval step [29], requirements with no counterpart in general-purpose RAG settings. These constraints are also shared by libraries, archives, and institutional repositories that hold large digitised collections but lack corpus-specific supervision or local GPU infrastructure.

In response, we propose TRACE (Training-free Retrieval with Agentic Corpus Exploration), a lightweight, training-free agentic RAG framework designed for accountable source discovery over historical corpora. This work is grounded in DECIDON,[1] a French National Research Agency project on political discourse circulation between parliamentary debates and the press during the French Third Republic, involving computer scientists, historians, and the Bibliothèque nationale de France (BnF). In historical and archival research, retrieval is not a hidden preprocessing step but a scholarly operation in its own right: a retrieved source must be discoverable, inspectable, and citable. TRACE is also intended to complement Corpusense, the IIIF-based tool developed within the Mezanno

[1] https://anr.fr/Projet-ANR-25-CE38-4063

CIKM '26, Rome, Italy

ACM ISBN 978-1-4503-XXXX-X/2026/11
https://doi.org/XXXXXXX.XXXXXXX

project for transforming digitised serial sources into structured data [1, 5, 23]. An open-source version of TRACE is also released [2].

Our contributions are threefold:

- We formulate historical archival retrieval as an applied, corpus-aware document retrieval problem under OCR degradation, heterogeneous source structure, and source-traceability constraints.
- We introduce TRACE, a training-free agentic retrieval framework combining corpus-targeted decomposition, multi-channel fused warm start, explicit retrieval verdicts, deferred re-evaluation, and count-grouped reranking.
- We evaluate TRACE on a 1,752-question French historical benchmark derived from BnF-digitised parliamentary and newspaper collections, showing consistent gains over sparse, dense, graph-based, and agentic RAG baselines.

2 Related Work

2.1 From Retrieval Effectiveness to Retrieval Accountability

Agentic RAG has emerged in two broad forms: training-based methods using reinforcement learning [3, 17, 21, 33] and training-free methods relying on framework design [4, 11, 25]. Both demonstrate meaningful retrieval gains, but share a critical limitation: they are evaluated end-to-end on generation metrics (Exact Match, F1, LLM-judged accuracy), leaving retrieval quality unexamined and conflating parametric knowledge with grounded evidence retrieval.[3] Graph-based RAG [12, 14, 15, 39] offers complementary structural indexing with consistent multi-hop gains, but introduces two constraints relevant here: graph construction is a form of lossy compression, and entity extraction degrades severely under OCR noise, causing structural errors that propagate into retrieval [39]. Historical archival corpora, which combine substantial OCR noise, heterogeneous sub-corpora, and strict source-traceability requirements, strain the assumptions of both paradigms simultaneously, motivating a retrieval-first framework that operates directly on noisy text without intermediate structural extraction.

2.2 RAG for Historical Corpora in Digital Humanities

RAG has been increasingly explored in digital humanities for querying large, heterogeneous, and OCR-noisy archival collections spanning newspaper archives [24, 31], parliamentary debates [28], classical texts [13], administrative records [19], and personal diary archives [22, 38]. A recurring finding is that corpus-specific properties, such as OCR artifacts, archaic vocabulary, and temporal structure, consistently dominate retrieval performance: named-entity injection at reranking partially mitigates OCR errors [31]; date-based metadata filtering substantially improves precision on chronologically dense collections [19]; and dense retrieval systematically fails on historical vocabulary, with sparse retrieval and reranking outperforming embedding-based approaches [35].

Alongside retrieval performance, a parallel concern has emerged around source traceability and human oversight [19, 28, 38]: in historical scholarship, identifying relevant sources is a scholarly objective in its own right, and retrieval must preserve an inspectable, citable path from query to evidence. These studies motivate a retrieval-first evaluation perspective in which source discovery, not only answer generation, becomes the object of measurement.

3 Methodology

3.1 Problem Setting and Architecture Overview

Let $\mathcal{D} = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K$ be a heterogeneous archival corpus partitioned into K sub-corpora that differ in genre, vocabulary, and document quality. Given a question q in the language of the corpus, the system must return an ordered list of document identifiers. The gold set $G_q \subseteq \mathcal{D}$ may span multiple sub-corpora and is not recoverable from the surface form of q. We evaluate with recall@k as the primary metric.

TRACE addresses the challenges identified above through a six-stage pipeline, as shown in Figure 1.

3.2 Planner: Corpus-Targeted Decomposition

The planner is a single LLM call that decomposes q into a set of M sub-questions $\{s_1, \ldots, s_M\}$, where $M \leq M_{\max}$. Each s_i is a tuple $(\text{text}_i, c_i, \text{entities}_i)$ where $c_i \in C$ identifies the target sub-corpus. The design follows the broader intuition of plan-and-execute and decomposition-based agentic retrieval [2, 3, 32]. The prompt enforces three invariants: (i) the sub-questions jointly cover the meaning of q without loss or addition; (ii) each s_i targets exactly one sub-corpus, so cross-corpus questions are fanned out into per-corpus sub-questions; (iii) entity names retain their original-language surface form to avoid OCR mismatch from translation.

Corpus targeting guides Phase 1 candidate allocation but does not restrict the agent loop: the agent may search across corpus boundaries when evidence suggests cross-corpus connections (§3.4).

3.3 Phase 1: Fused Warm Start

For each sub-question s_i, Phase 1 produces an initial candidate pool by fusing four complementary retrieval channels via reciprocal-rank fusion (RRF, k=60) [8]. Each channel independently scores chunks and *max-pools* per parent document: a document with N chunks contributes exactly once per channel—its best chunk score—preventing any single long document from dominating the pool through sheer volume of passages. This hybrid design follows recent findings that sparse, dense, metadata-aware, and reranking strategies often compensate for different failure modes in historical corpora [19, 35].

Sparse lexical (BM25) [30] retrieves over tokenised chunks and is effective for proper nouns and domain terms that survive OCR.

Dense semantic uses cosine similarity over pre-computed chunk embeddings, capturing paraphrases and conceptual matches that lexical search misses.

Temporal uses date-based retrieval with exponential decay scoring by temporal distance from the target date, reflecting the importance of chronological metadata in historical retrieval settings [19].

[2] https://github.com/Kepler1908/TRACE

[3] Self-RAG [3] and RAGentA [4] do report recall, but as one metric among several.

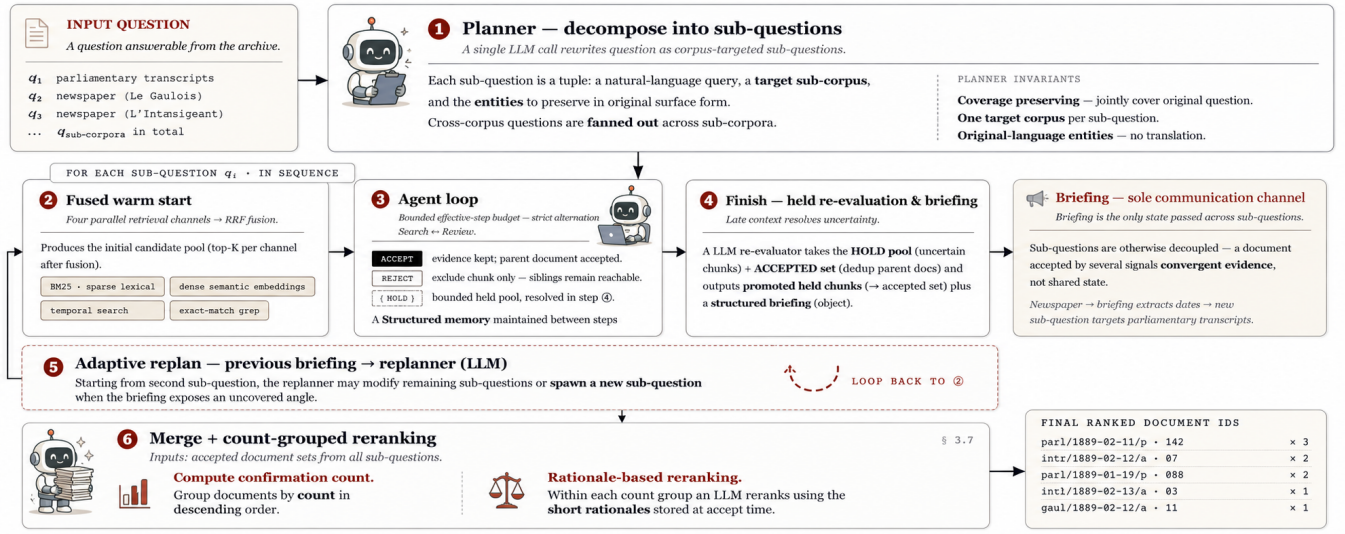

Figure 1: Overview of the TRACE pipeline.

Exact match uses substring search over raw chunk text, targeting rare proper nouns and OCR-resistant spellings that dense models may not encode faithfully.

The complementarity of these channels provides implicit robustness to OCR noise without explicit preprocessing: BM25 handles intact tokens, exact match handles partial substrings, and dense retrieval captures semantic paraphrases of corrupted text. In an imbalanced corpus, however, applying a global result cap systematically truncates minority sub-corpus documents. We address this by pre-computing the set of document IDs belonging to the target sub-corpus and passing these as *candidate constraints* to each channel, so the cap applies *within* the target sub-corpus rather than across the full collection.

3.4 Phase 2: Agent Loop

The warm-start pool seeds a bounded agent loop with an *effective step budget*. At each step, the agent either issues a new search query through one of the available channels or reviews a pending batch of retrieved chunks. Search and review strictly alternate: after each search returns candidates, the agent must review them before issuing another query, ensuring that each retrieval is immediately exploited and subsequent queries can build on newly discovered evidence. This design is related to ReAct-style alternation between reasoning and action [34], but adapts it to source discovery by making retrieval verdicts explicit. Every reviewed chunk receives one of three verdicts:

ACCEPT The chunk provides evidence toward answering q. The agent records a short free-text rationale (≤120 characters, reused at reranking time), and the parent document is deduplicated from subsequent retrieval.

REJECT The chunk is off-topic. Only the chunk—not the parent document—is excluded from future retrieval, preserving access to sibling passages of the same long document.

HOLD The chunk is possibly relevant but uncertain given current context. It enters a bounded *held pool* for deferred re-evaluation (§3.5).

The agent is always presented with both q and s_i: the sub-question focuses the search direction, but relevance is defined with respect to the original question, ensuring that cross-cutting evidence is not discarded by an overly narrow interpretation of the current sub-question.

Temporal search receives special treatment: the agent's specified date window is silently expanded by one day on each boundary, compensating for indeterminacy in date-boundary decisions without exposing the expansion to the agent. Exposing it would risk anchoring subsequent temporal reasoning on artificially widened boundaries. Results from date expansion are conservatively auto-held rather than accepted, entering the deferred pool for later re-evaluation.

To prevent circular exploration, the agent maintains a five-slot structured memory: prior findings, question analysis, evidence gathered, gaps remaining, and next steps. This memory is updated by *replacement* at every step. Replacement semantics force the agent to synthesise rather than accumulate, keeping the memory compact and current. Alongside this, the system tracks an *effective step counter* that increments only on productive actions. Duplicate queries, parse failures, and empty results are skipped with informative feedback, preserving the step budget for genuinely new exploration; two consecutive identical search attempts trigger forced termination to prevent degenerate loops.

Each agent step is a single-shot LLM call, receiving a system prompt and a constructed user prompt in JSON mode. The agent retains no memory of prior API calls; all state is reconstructed from the structured memory and a windowed action log containing the last eight entries. The output is a JSON object (`thought`, `action`, `args`, `memory_update`) parsed without tool-calling APIs.

This design eliminates context pollution from accumulated conversation history, makes calls stateless and parallelisable across sub-questions, and renders the system provider-agnostic: any LLM supporting JSON-mode output is sufficient, with no dependency on proprietary function-calling interfaces.

3.5 Finish, Briefing, and Replan

When the agent terminates, a *held re-evaluation* call presents the full accepted set, the stage memory, and summaries of all held chunks to the LLM, which promotes any held chunk now supported by the accumulated evidence. This is the payoff of the three-way verdict: early uncertainty is resolved against late context rather than collapsed into a premature binary decision at first encounter.

A *briefing* is then synthesised as a structured object containing a narrative summary, confirmed facts, confirmed dates, confirmed entities, remaining gaps, and search hints. It serves as the sole communication channel between sub-questions. This enables later sub-questions to start from the evidence frontier rather than searching blind. Before each non-initial sub-question s_i (i>1), a replanner receives the previous briefing and the remaining plan, and may modify remaining sub-questions or spawn new ones when the briefing reveals gaps not covered by the original decomposition. The total number of sub-questions is capped at $M_{\max}$ to bound cost. This transforms the retrieval plan from a static script into an adaptive investigation while keeping sub-question processing auditable.

3.6 Merge and Count-Grouped Reranking

After all sub-questions complete, their accepted sets $\{A_1, \ldots, A_M\}$ are merged. Because sub-questions are structurally decoupled and share no accepted sets during their respective agent loops, a document accepted by multiple sub-questions represents convergent evidence from separate investigative paths rather than confirmation bias from shared state. For each document $d \in \bigcup_i A_i$ we compute a *confirmation count* $\mathrm{cnt}(d) = |\{i : d \in A_i\}|$, and the final ranking proceeds in two stages. First, documents are partitioned by $\mathrm{cnt}(d)$ and groups are placed in decreasing count order. Within each group of size ≥ 2, an LLM ranks the documents using the concatenation of the agent's own acceptance rationales—the ≤120-character justifications recorded during the agent loop—rather than re-reading full document text. Singleton groups pass through without an LLM call, so the reranker's cost is proportional to the number of ties rather than the total candidate count.

4 Experimental Setup

4.1 Corpus and Questions

We use HistoriQA-ThirdRepublic [27] as our evaluation dataset. It consists of a text corpus and a question-answer dataset built around the same type of historical materials that motivate the DECIDON use case: parliamentary debates and newspapers from the French Third Republic. The benchmark therefore serves both as an evaluation resource and as a controlled proxy for the broader applied problem of accountable source discovery over BnF-derived digitised collections.

The corpus consists of 3,386 French-language documents from the French Third Republic period (1887), drawn from three heterogeneous sub-corpora: parliamentary debate transcripts from the *Journal Officiel de la République française* and two Parisian daily newspapers (*Le Gaulois*, *L'Intransigeant*) representing distinct editorial positions on the political events of the period. The documents are derived from OCR-digitised historical collections held by the BnF, reflecting the type of large-scale digitised material encountered in national library and archival infrastructures. They contain substantial OCR errors, irregular layouts, and genre-specific vocabulary, making the corpus a realistic testbed for accountable retrieval in digital heritage settings. Documents are split into 5,664 sentence-boundary chunks with a maximum of 1,000 whitespace tokens and no overlap.

The evaluation set comprises 1,752 questions with gold document sets, constructed following the methodology of Pellet et al. [27] in collaboration with a domain historian, who validated the factual and contextual coherence of the generated questions. Questions span four types reflecting patterns of historical inquiry: single-hop (**SH**, n=889), where each question targets a single source document; multi-hop generic (**MH-Generic**, n=529), requiring cross-source synthesis across parliamentary and press sources; bridge-entity (**MH-BridgeEnt**, n=142), where a shared named entity connects documents from different sub-corpora; and comparative (**MH-Comp**, n=192), contrasting viewpoints across heterogeneous sources. Gold sets may span multiple sub-corpora, making corpus-aware retrieval essential. Prior evaluation on this corpus demonstrates the severity of the retrieval challenge: standard dense retrieval achieves a Recall@3 of only 14.3% on cross-newspaper queries and 47.8% on cross-domain queries [27].

4.2 Baselines and Metrics

We conduct three sets of experiments. **(1) Main results** evaluate TRACE under two backbone configurations: DeepSeek-V3.2 [9], the earlier model used during system development, and DeepSeek-V4-Flash [10], the configuration adopted for the remaining experiments. **(2) Baseline comparison** compares TRACE against six systems across three retrieval paradigms: basic retrieval (BM25 and dense retrieval), graph-based RAG (HippoRAG 2 [15] and LinearRAG [39]), and agentic RAG (A-RAG [11] and MA-RAG [25]). These baselines are selected for the distinctiveness of their design or strong reported performance on multi-hop retrieval benchmarks. **(3) Cumulative ablation** progressively adds components to a baseline agent loop to quantify the contribution of each mechanism. Experiments (2) and (3) use DeepSeek-V4-Flash and Qwen3-Embedding-8B [37] throughout.

All baseline systems were adapted with necessary modifications, such as prompt translation and model replacement for SpaCy [16], to operate correctly on the French corpus. All systems share the same chunking, embedding model, and LLM backbone wherever applicable, ensuring that differences reflect retrieval architecture rather than corpus preprocessing or model choice.

For a question with gold set G_q and ranked result list π, we report:

$$\mathrm{R@}k = \frac{|G_q \cap \pi_{1:k}|}{|G_q|}, \qquad \mathrm{MRR}(q) = \frac{1}{|G_q|} \sum_{g \in G_q} \frac{1}{\mathrm{rank}_\pi(g)},$$

where $\mathrm{rank}_\pi(g) = \infty$ if $g \notin \pi$. This multi-gold MRR averages the reciprocal rank over *all* gold documents, penalising systems that

surface only the easiest gold document while burying the rest. For agentic systems, we additionally report precision $P_{\text{acc}} = |A_q \cap G_q| / |A_q|$ and recall $R_{\text{acc}} = |A_q \cap G_q| / |G_q|$ over the accepted set A_q, measuring the quality of accept/reject decisions independently of the final ranking. In the main experiments, we also report the average number of accepted documents for each question type.

5 Results

5.1 Main Results

Table 1 reports per-type retrieval performance on the full 1,752-question evaluation set. The two backbone models exhibit a consistent precision-recall trade-off: DeepSeek-V3.2 achieves higher overall recall by accepting more documents on average, while DeepSeek-V4-Flash yields substantially higher P_{acc} with a more selective acceptance strategy. Both models show the same difficulty gradient across question types: single-hop questions are near-ceiling, multi-hop generic questions maintain high recall but lower MRR, and bridge-entity and comparative questions remain hardest as they require cross-corpus evidence chains. Across all types, R_{acc} closely tracks R@10, indicating that the performance bottleneck lies in *finding and accepting* gold documents during the agent loop rather than in the final reranking step.

The more conservative acceptance behaviour of DeepSeek-V4-Flash should be interpreted as a different operating point rather than a simple degradation: the agent declines to accept marginally relevant documents, improving accepted-set precision while modestly reducing accepted-set recall. This matters in digital-archive workflows, where every accepted document may become part of a human reading queue. A configuration that returns a smaller, more precise accepted set can therefore be preferable even when raw recall is slightly lower.

In our hosted inference setup, DeepSeek-V4-Flash consumed a total of 177.6 M input tokens and 39.7 M output tokens, costing approximately $36 at the API prices used for the experiment. DeepSeek-V3.2 cost approximately $93, or 158% more, driven by higher per-token pricing and 1.8× greater input consumption from deeper agent exploration. At approximately $0.02 per question, the V4-Flash configuration makes large-scale corpus exploration economically feasible for humanities laboratories operating under tight resource constraints, while remaining well below the practical cost of maintaining equivalent local GPU deployment for occasional or project-based use. Given these deployment advantages, we adopt DeepSeek-V4-Flash as the backbone for practical use.

5.2 Comparison with Baselines

Table 2 compares TRACE against representative baselines across three retrieval paradigms.

On basic retrieval, dense retrieval performs well on single-hop questions but degrades substantially on multi-hop types, a pattern consistent with what Pellet et al. [27] observed on this corpus. Sparse retrieval (BM25) is competitive on single-hop questions but performs poorly on comparative questions, where relevant evidence is dispersed across heterogeneous sources and lexical overlap between the query and the evidence may be limited.

Graph-based systems yield a more surprising result: both HippoRAG 2 and LinearRAG perform at levels comparable to or below dense retrieval across most question types. We hypothesise that OCR noise in the corpus degrades entity extraction and relation linking during graph construction, causing graph-based methods to lose part of their structural advantage. This is consistent with the concern raised in Section 2: graph-based indexing can amplify rather than correct corpus quality issues when the extracted structure is unreliable.

TRACE achieves consistent and substantial gains across all question types, with the margin widening on multi-hop questions that require cross-corpus evidence chains.

The two agentic baselines, A-RAG and MA-RAG, do not produce explicit ranked document lists, instead generating answers from their full retrieval trajectory. R@k and MRR are therefore not applicable. We use P_{acc} and R_{acc} as the basis for comparison, treating their accumulated retrieved documents as the accepted set. This is a favourable proxy: it gives the baselines credit for any relevant document encountered during their trajectory, regardless of whether the document was ultimately used in a generated answer. Even under this generous interpretation, TRACE achieves substantially higher precision and recall simultaneously, indicating that the advantage stems from the quality of individual retrieval decisions rather than from a larger candidate pool.

The purpose of this comparison is not simply to establish that TRACE outperforms existing systems. More importantly, it surfaces a structural blind spot in how RAG systems are typically evaluated: generation-centric metrics obscure the quality of the underlying retrieval process, and systems optimised for open-domain QA do not naturally transfer to domain-specific deployment. Effective retrieval in historical research requires targeted design choices.

5.3 Cumulative Ablation Study

Table 3 reports a cumulative ablation in which components are added one at a time to a baseline agent loop. Because the mechanisms in TRACE are designed to interact—decomposition produces the sub-questions that memory and replanning operate on, while the warm start seeds the pool that the agent loop refines—we adopt a cumulative rather than leave-one-out design, which better reflects the system's intended operating mode.

Sub-question decomposition yields the largest gain (+8.9%), dramatically improving multi-hop types while leaving single-hop performance largely unchanged. Memory and replanning (+2.4%) primarily benefits bridge-entity questions by enabling evidence to accumulate across sub-question boundaries. The Phase 1 warm start (+5.0%) stabilises all question types by seeding the agent with a high-quality initial candidate pool. The hold-and-re-evaluation mechanism contributes modestly in aggregate (+0.5%), but acts as a conservative safety layer: its value is concentrated in marginal cases where a premature rejection would permanently foreclose a document whose relevance only becomes clear under late-accumulated context.

6 Discussion, Deployment, and Future Work

The results show that accountable source discovery over noisy historical archives benefits less from heavier indexing or model training than from corpus-aware orchestration. TRACE combines four design choices that proved particularly important in

Table 1: Main retrieval results on the full evaluation set (N=1,752). #RET is the mean number of accepted documents.

Type	R@3	R@5	R@10	MRR	P_{acc}	R_{acc}	#RET
DeepSeek-V3.2 (0324)							
SH	0.872	0.904	0.922	0.797	0.246	0.933	8.1
MH – Generic	0.684	0.808	0.892	0.532	0.250	0.932	11.0
MH – BridgeEnt	0.644	0.761	0.820	0.508	0.276	0.838	8.5
MH – Comparative	0.484	0.596	0.727	0.383	0.140	0.846	21.1
All	0.754	**0.830**	**0.884**	0.648	0.238	**0.915**	10.4
DeepSeek-V4-Flash							
SH	0.852	0.881	0.893	0.793	0.452	0.902	4.6
MH – Generic	0.739	0.832	0.889	0.560	0.349	0.906	8.5
MH – BridgeEnt	0.648	0.725	0.750	0.472	0.348	0.754	5.7
MH – Comparative	0.500	0.578	0.674	0.396	0.275	0.789	14.9
All	**0.763**	0.820	0.856	**0.653**	**0.393**	0.879	**7.0**

Table 2: Baseline comparison on HistoriQA-ThirdRepublic. Panel (a) reports R@10 and MRR for all systems. Panel (b) reports accepted-set precision and recall for agentic systems.

(a) R@10 and MRR per question type (all systems)

System	SH		MH – Generic		MH – BridgeEnt		MH – Comp		All	
	R@10	MRR	R@10	MRR	R@10	MRR	R@10	MRR	R@10	MRR
Basic retrieval										
BM25	0.760	0.598	0.571	0.370	0.556	0.370	0.148	0.084	0.620	0.454
Dense retrieval	0.865	0.712	0.621	0.380	0.606	0.391	0.250	0.105	0.703	0.519
Graph-based RAG										
HippoRAG 2	0.867	0.575	0.629	0.354	0.602	0.365	0.273	0.105	0.709	0.440
LinearRAG	0.864	0.687	0.614	0.364	0.616	0.393	0.245	0.098	0.701	0.501
Agentic RAG										
A-RAG	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
MA-RAG	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
TRACE (ours)	**0.893**	**0.793**	**0.889**	**0.560**	**0.750**	**0.472**	**0.674**	**0.396**	**0.856**	**0.653**

(b) Agent decision quality (P_{acc} / R_{acc}) per question type (agentic systems only)

System	SH		MH – Generic		MH – BridgeEnt		MH – Comp		All	
	P_{acc}	R_{acc}	P_{acc}	R_{acc}	P_{acc}	R_{acc}	P_{acc}	R_{acc}	P_{acc}	R_{acc}
A-RAG	0.048	0.595	0.072	0.469	0.082	0.493	0.016	0.177	0.054	0.503
MA-RAG	0.055	0.774	0.118	0.639	0.130	0.627	0.045	0.271	0.079	0.666
TRACE (ours)	**0.452**	**0.902**	**0.349**	**0.906**	**0.348**	**0.754**	**0.275**	**0.789**	**0.393**	**0.879**

this setting: decomposition into corpus-targeted sub-questions, multi-channel fused retrieval, explicit accept/reject/hold decisions, and final ranking based on cross-sub-question confirmation. On HistoriQA-ThirdRepublic, this design achieves R@10 = 0.856 and MRR = 0.653, outperforming sparse, dense, graph-based, and agentic RAG baselines while remaining training-free.

The DECIDON use case also highlights why retrieval should be treated as an applied scholarly task rather than as a hidden preprocessing step for answer generation. In digital-heritage settings, users need to identify relevant sources, understand why they were retrieved, and preserve a traceable path from query to evidence. TRACE is currently deployed as an internal prototype within the DECIDON project and is accessible to 24 researchers from six partner institutions: EPITA, EHESS-CRH, Inria, the Bibliothèque nationale de France, the École nationale des chartes, and LARHRA. As of submission, the prototype has been accessible to project researchers since March 2026 and has supported several hundred exploratory research queries over parliamentary debates and the press.

The prototype supports four historical case studies: the politicisation of nature, the Colonial Party, social and tax legislation, and anti-parliamentarianism during the French Third Republic. In these cases, historians use TRACE to locate relevant passages, compare documentary contexts, and follow the circulation of actors,

Table 3: Cumulative ablation on the full evaluation set (N=1,752). Each level adds one component to the previous level. Δ is the change in overall R@10 from the previous level.

		SH		MH – Generic		MH – BridgeEnt		MH – Comp		All	
Lv	**Added component**	R@10	MRR	R@10	MRR	R@10	MRR	R@10	MRR	R@10	Δ
0	Baseline (agent only)	0.813	0.713	0.647	0.435	0.352	0.247	0.466	0.284	0.688	—
1	+ Decomposition	0.790	0.699	0.849	0.539	0.637	0.413	0.625	0.368	0.777	+8.9%
2	+ Memory & replan	0.807	0.707	0.846	0.536	0.775	0.455	0.669	0.371	0.801	+2.4%
3	+ Phase 1 warm start	0.885	0.785	0.882	0.559	0.736	0.471	0.695	0.411	0.851	+5.0%
4	+ Hold & re-eval (full)	0.893	0.793	0.889	0.560	0.750	0.472	0.674	0.396	0.856	+0.5%

arguments, vocabularies, and framings over time. The system remains a research service rather than a public-facing BnF service, but the planned integration with Corpusense would support a broader BnF-facing workflow in which TRACE helps identify and document relevant sources, while Corpusense supports the downstream transformation of selected IIIF collections or serial documents into structured data [1, 5, 23].

Several limitations remain. First, the planner occasionally generates a single sub-question for inherently multi-hop queries, preventing the replan mechanism from firing. Second, the benchmark covers a single historical year, 1887; generalisation to other periods, languages, and archival genres remains to be tested. Third, some prompt-level assumptions, such as the relation between parliamentary debates and newspapers, are corpus-specific and would require adjustment for other documentary settings. Finally, preliminary local-inference tests on an RTX PRO 6000 workstation showed that smaller models still struggle with multi-step planning and structured-output stability: a gpt-oss-20b model [26] achieved only 17.5% task completion, while a fine-tuned Qwen3.5-27B model [18] reached 40% on single-hop queries but incurred a 26.7% JSON parse error rate. This suggests that local deployment will require either larger institution-grade infrastructure or a protocol-optimised version of TRACE.

Future work will focus on making TRACE a mixed-initiative source discovery system. Rather than placing historians only at the end of the pipeline, we plan to test lightweight checkpoints where expert users can validate or revise the planner's decomposition, inspect held documents before final rejection, and adjust the final accepted set. Additional deployment work will measure latency, token consumption, source reuse, and expert reading burden across larger samples, enabling a more precise evaluation of the trade-off between retrieval accuracy, computational cost, and human workload.

Acknowledgments

This work was carried out in the context of the DECIDON project, funded by the French National Research Agency under grant ANR-25-CE38-4063. The experiments build on digitised historical collections from the Bibliothèque nationale de France. Florian Cafiero was supported by the PSL Research University's Major Research Program CultureLab, implemented by the ANR (reference ANR-10-IDEX-0001).

GenAI Usage Disclosure

Generative AI tools were used to assist with language editing and code completion during the preparation of this work. All scientific claims, experimental design choices, analyses, reported results, and final text were produced, reviewed, and validated by the authors, who remain fully accountable for the content of the paper.

References

[1] Nathalie Abadie, Marie Carlin, Edwin Carlinet, Joseph Chazalon, Pascal Cristofoli, et al. 2026. Mezanno : des sources sérielles aux données structurées pour les humanités numériques. Démonstrations de la conférence EGC 2026. https://hal.science/hal-05444727

[2] Paul J. L. Ammann, Jonas Golde, and Alan Akbik. 2025. Question Decomposition for Retrieval-Augmented Generation. arXiv:2507.00355. https://arxiv.org/abs/2507.00355

[3] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. arXiv:2310.11511. doi:10.48550/arXiv.2310.11511

[4] Ines Besrour, Jingbo He, Tobias Schreieder, and Michael Färber. 2025. RAGentA: Multi-Agent Retrieval-Augmented Generation for Attributed Question Answering. arXiv:2506.16988. doi:10.48550/arXiv.2506.16988

[5] Bibliothèque nationale de France. 2026. Mezanno : libérer les données des archives numérisées. Chroniques de la BnF, no. 105, avril–juillet 2026. https://www.bnf.fr/fr/mezanno-liberer-les-donnees-des-archives-numerisees-echos-de-recherche

[6] Francis X. Blouin Jr. and William G. Rosenberg. 2011. Authoritative History and Authoritative Archives. In *Processing the Past: Contesting Authorities in History and the Archives*, Francis X. Blouin Jr. and William G. Rosenberg (Eds.). Oxford University Press, Oxford, UK. doi:10.1093/acprof:oso/9780199740543.003.0002

[7] Jonathan Bourne. 2026. Reading the unreadable: creating a dataset of 19th century English newspapers using image-to-text language models. *Digital Scholarship in the Humanities* 41, 1 (2026), 22–40. doi:10.1093/llc/fqaf151

[8] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 758–759. doi:10.1145/1571941.1572114

[9] DeepSeek-AI. 2025. DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models. arXiv:2512.02556. doi:10.48550/arXiv.2512.02556

[10] DeepSeek-AI. 2026. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence. Technical report. https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro/blob/main/DeepSeek_V4.pdf

[11] Mingxuan Du, Benfeng Xu, Chiwei Zhu, Shaohan Wang, Pengyu Wang, Xiaorui Wang, and Zhendong Mao. 2026. A-RAG: Scaling Agentic Retrieval-Augmented Generation via Hierarchical Retrieval Interfaces. arXiv:2602.03442. doi:10.48550/arXiv.2602.03442

[12] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. arXiv:2404.16130. doi:10.48550/arXiv.2404.16130

[13] Yang Fan, Zhang Qi, Xing Wenqian, Liu Chang, and Liu Liu. 2025. Research on Graph-Retrieval Augmented Generation Based on Historical Text Knowledge Graphs. arXiv:2506.15241. doi:10.48550/arXiv.2506.15241

[14] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2025. LightRAG: Simple and Fast Retrieval-Augmented Generation. arXiv:2410.05779. doi:10.48550/arXiv.2410.05779

[15] Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. From RAG to Memory: Non-Parametric Continual Learning for Large Language Models. arXiv:2502.14802. doi:10.48550/arXiv.2502.14802

[16] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. Software. doi:10.5281/zenodo.1212303

[17] Yulong Hui, Chao Chen, Zhihang Fu, Yihao Liu, Jieping Ye, and Huanchen Zhang. 2026. Interact-RAG: Reason and Interact with the Corpus, Beyond Black-Box Retrieval. arXiv:2510.27566. doi:10.48550/arXiv.2510.27566

[18] Jackrong. 2026. Jackrong-llm-finetuning-guide: An Educational LLM Fine-Tuning Pipeline. GitHub repository. https://github.com/Jackrong/Jackrong-llm-finetuning-guide

[19] Jeong Ha Lee, Ghazanfar Ali, and Jae-In Hwang. 2025. A Retrieval-Augmented Generation System for Accurate and Contextual Historical Analysis: AI-Agent for the Annals of the Joseon Dynasty. *Computer Animation and Virtual Worlds* 36, 4 (2025), e70048. doi:10.1002/cav.70048

[20] 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*, Vol. 33. 9459–9474. https://arxiv.org/abs/2005.11401

[21] Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic Search-Enhanced Large Reasoning Models. arXiv:2501.05366. doi:10.48550/arXiv.2501.05366

[22] Claire Lin, Bo-Han Feng, Xuanjun Chen, Te-Lun Yang, Hung-yi Lee, and Jyh-Shing Roger Jang. 2025. A Preliminary Study of RAG for Taiwanese Historical Archives. arXiv:2511.07445. doi:10.48550/arXiv.2511.07445

[23] Mezanno Project. 2026. Corpusense. Project website. https://mezanno.xyz/corpusense/ Accessed 2026-05-13.

[24] Anthony Mudet and Souhail Bakkali. 2025. Hybrid Retrieval-Augmented Generation for Robust Multilingual Document Question Answering. arXiv:2512.12694. https://arxiv.org/abs/2512.12694

[25] Thang Nguyen, Peter Chin, and Yu-Wing Tai. 2025. MA-RAG: Multi-Agent Retrieval-Augmented Generation via Collaborative Chain-of-Thought Reasoning. arXiv:2505.20096. doi:10.48550/arXiv.2505.20096

[26] OpenAI. 2025. gpt-oss-120b and gpt-oss-20b Model Card. arXiv:2508.10925. https://arxiv.org/abs/2508.10925

[27] Aurélien Pellet, Marie Puren, and Julien Perez. 2026. HistoriQA-ThirdRepublic: Multi-Hop Question Answering Corpus for Historical Research, Parliamentary Debates from the French Third Republic (1870–1940). In *Proceedings of the Language Resources and Evaluation Conference*. ELRA Language Resources Association, Palma de Mallorca, Spain. https://hal.science/hal-05438255

[28] Julien Perez, Aurélien Pellet, and Marie Puren. 2025. Évaluation automatique du retour à la source dans un contexte historique long et bruité. Application aux débats parlementaires de la Troisième République française. In *Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, Frédéric Bechet, Adrian-Gabriel Chifu, Karen Pinel-sauvagnat, Benoit Favre, Eliot Maes, and Diana Nurbakova (Eds.). ATALA and ARIA, Marseille, France, 138–150. https://aclanthology.org/2025.jeptalnrecital-evalllm.12/

[29] Marie Puren, Donghan Bian, Aurélien Pellet, Julien Perez, and Florian Cafiero. 2026. Aligner méthode historique et RAG : transformer un assistant conversationnel en chaîne de preuve auditable et discutable. HAL preprint. https://hal.science/hal-05460105

[30] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* 3, 4 (2009), 333–389. doi:10.1561/1500000019

[31] The Trung Tran, Carlos-Emiliano González-Gallardo, and Antoine Doucet. 2025. Retrieval Augmented Generation for Historical Newspapers. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3677389.3702542

[32] Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. arXiv:2305.04091. doi:10.48550/arXiv.2305.04091

[33] Tianle Xia, Ming Xu, Lingxiang Hu, Yiding Sun, Wenwei Li, Linfang Shang, Liqun Liu, Peng Shu, Huan Yu, and Jie Jiang. 2026. Search-P1: Path-Centric Reward Shaping for Stable and Efficient Agentic RAG Training. arXiv:2602.22576. doi:10.48550/arXiv.2602.22576

[34] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629. doi:10.48550/arXiv.2210.03629

[35] Yongan Yu, Xianda Du, Qingchen Hu, Jiahao Liang, Jingwei Ni, Dan Qiang, Kaiyu Huang, Grant McKenzie, Renee Sieber, and Fengran Mo. 2025. WeatherArchive-Bench: Benchmarking Retrieval-Augmented Reasoning for Historical Weather Archives. arXiv:2510.05336. doi:10.48550/arXiv.2510.05336

[36] Junyuan Zhang, Qintong Zhang, Bin Wang, Linke Ouyang, Zichen Wen, Ying Li, Ka-Ho Chow, Conghui He, and Wentao Zhang. 2025. OCR Hinders RAG: Evaluating the Cascading Impact of OCR on Retrieval-Augmented Generation. arXiv:2412.02592. doi:10.48550/arXiv.2412.02592

[37] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. arXiv:2506.05176. doi:10.48550/arXiv.2506.05176

[38] Jing Zhou, Li Si, and Wenjun Hou. 2025. Humanities-in-the-Loop: Using Close Reading as a Method for Retrieval-Augmented Generation (RAG). *Proceedings of the Association for Information Science and Technology* 62, 1 (2025), 1185–1189. doi:10.1002/pra2.1355

[39] Luyao Zhuang, Shengyuan Chen, Yilin Xiao, Huachi Zhou, Yujing Zhang, Hao Chen, Qinggang Zhang, and Xiao Huang. 2025. LinearRAG: Linear Graph Retrieval Augmented Generation on Large-scale Corpora. arXiv:2510.10114. doi:10.48550/arXiv.2510.10114