Giving Agents a Map: Structured Indexes for Fact Coverage in Complex Document Collections

Yuyingzi Yang* Sijia Pan* Guangju Wang
{yangyuyingzi,pansijia,wangguangju}@benzhengluoji.com

Abstract

Analytical decision-making over large, heterogeneous document collections depends on comprehensive evidence gathering and synthesis. To support this process at scale, we propose procedure-index, which compiles a document collection into a navigation surface organized around a fixed analytical procedure. It exposes typed modules, route documents, cross-module views, and provenance-linked evidence items, which an LLM agent traverses with read-only tools. This study compares procedure-index with whole-document embedding search on 50 due-diligence cases over a frozen snapshot of 29,688 documents. Under the main deployed protocols, procedure-index records higher final-answer fact recall on all four primary metrics, with all four main-comparison company-cluster tests significant, at approximately 6.3 times the recorded model-API token usage. In a sequential interface ablation, the largest observed adjacent near-hit increment occurs when the navigation-document package—route organization plus short evidence cues—is added to hierarchical rendered evidence. Across 10 shared nominal evidence-collection caps, natural-stop procedure-index remains descriptively higher on strict and near-hit recall. An embedding-only forced-continuation follow-up increases realized acquisition effort; its observed pooled recall estimates vary non-monotonically across the budget grid and remain below the natural-stop procedure-index reference at the high-budget endpoint. Selected cases are consistent with higher coverage when target facts are dispersed across evidence regions and with a boundary where a few query-similar documents concentrate the answer. These results motivate treating agent-visible corpus structure and stopping policy as design dimensions for high-recall answer coverage in due diligence.

1 Introduction

LLM agents are increasingly expected to gather evidence actively: they can reformulate a query, inspect a source, follow a lead, and decide what to examine next. Hierarchical corpus navigation supports this behavior by exposing document structure and letting agents choose what to inspect (Sun et al., 2026; Dai et al., 2026; Du et al., 2026; Zhao and Yang, 2026). Heterogeneous company due diligence makes the choice of access interface concrete: final answers must synthesize multiple query-scoped facts that may be dispersed across documents, and omitting one can alter conclusions about ownership, financing, or litigation.

This study compares procedure-index with a whole-document embedding comparator that presents the agent with a rank-truncated candidate window. The agent issues a query, receives the top-ranked documents, and repeats, but this interface does not reveal which corpus regions exist or which routes remain

*Equal contribution.

unexplored. By contrast, a procedure-structured interface can expose analytical regions and routes before item-level inspection. We therefore ask whether the complete procedure-structured interface differs from this comparator in final-answer fact coverage, which cumulative interface packages account for observed increments, and how the contrast behaves when nominal evidence-access budgets and stopping policies change.

Procedure-index is a procedure-structured instance of agent-visible hierarchical navigation, traversed with read-only tools. It progressively exposes the collection from source hierarchy and typed modules through route documents to provenance-linked evidence items. The map metaphor denotes this corpus-level orientation before item-level inspection. We use this interface to study final-answer target-fact coverage on a company due-diligence benchmark, comparing it with a whole-document embedding interface while holding the target set, underlying source environment, answer-model name, and recognizer fixed. Strict recall counts fully expressed targets; near-hit recall also credits non-contradicted partial coverage. The outcome measures citation-agnostic target-fact expression in final answers rather than verified support from opened evidence.

In the main comparison, procedure-index has higher recall on all four primary metrics (planned exact company-cluster $p \leq 0.006$), with approximately 6.3 times the recorded model-API token usage. The sequential interface ablation locates the largest observed adjacent near-hit increment when the navigation-document package is added to hierarchical rendered evidence (+0.133 critical near-hit recall), rather than in rendered evidence alone. An initial \$0.40 constrained-budget comparison serves as a pilot operating point: its overall near-hit contrast is positive but suggestive (+0.044, exact company-cluster $p = 0.059$), while the other three paired contrasts are not significant. The subsequent multi-budget analyses provide broader evidence that the observed advantage is not confined to a single nominal cap. Across 10 shared natural-stop caps, procedure-index remains descriptively higher than embedding on strict and near-hit recall, but this is a complete-interface comparison rather than an equal-realized-cost claim. An embedding-only forced-continuation follow-up separates acquisition effort from ordinary stopping: realized spending, search calls, and fetch calls increase, and recall rises overall but non-monotonically across the grid. At \$2.00, forced embedding remains below the natural-stop procedure-index reference. Because there is no matched forced procedure-index arm, this follow-up supports the interpretation that budget and stopping policy matter; it does not identify a structure-only effect. Selected cases (Section 6) are consistent with a dispersion-based account and its boundary: direct ranking can be more effective when a few query-similar articles already concentrate the answer.

The evaluated treatment is the deployed interface as a bundle—representation, artifacts, tools, prompts, budget, and stopping policy—so the main contrast is a paired system-level comparison. The sequential interface ablation and budget analyses qualify this contrast but do not isolate every factor. This preprint discloses the measurement contract and translated, partially redacted prompt contracts needed to interpret the reported aggregates.

This paper makes two contributions:

- We report a paired, system-level evaluation of agent-visible navigation for heterogeneous due diligence, instantiated in the procedure-index interface. Under the deployed protocols, the complete interface records higher final-answer fact recall than whole-document embedding search on all four primary metrics. A sequential interface ablation over hierarchical rendered evidence, navigation documents, and skill guidance locates the largest adjacent near-hit increment when the navigation-document package is added, a package that combines route organization with short evidence cues. A 10-point natural-stop budget sensitivity analysis retains a descriptive procedure-index advantage, while an embedding-only forced-continuation follow-up shows that increasing realized acquisition effort does not produce a stable monotonic recall response or eliminate the high-budget descriptive gap. Selected traces further suggest that the advantage concentrates where target facts are dispersed across evidence regions, with a boundary when a few query-similar documents already concentrate the answer.
- We specify and disclose a fact-level measurement contract for final-answer coverage: 3,321 frozen, query-scoped target facts with priority tiers, scored by a closed-world recognizer that quotes answer spans and separates coverage, contradiction, and unscored decisions, with company-cluster exact tests. This contract adapts nugget-style fact coverage to a frozen, access-controlled evidence setting.

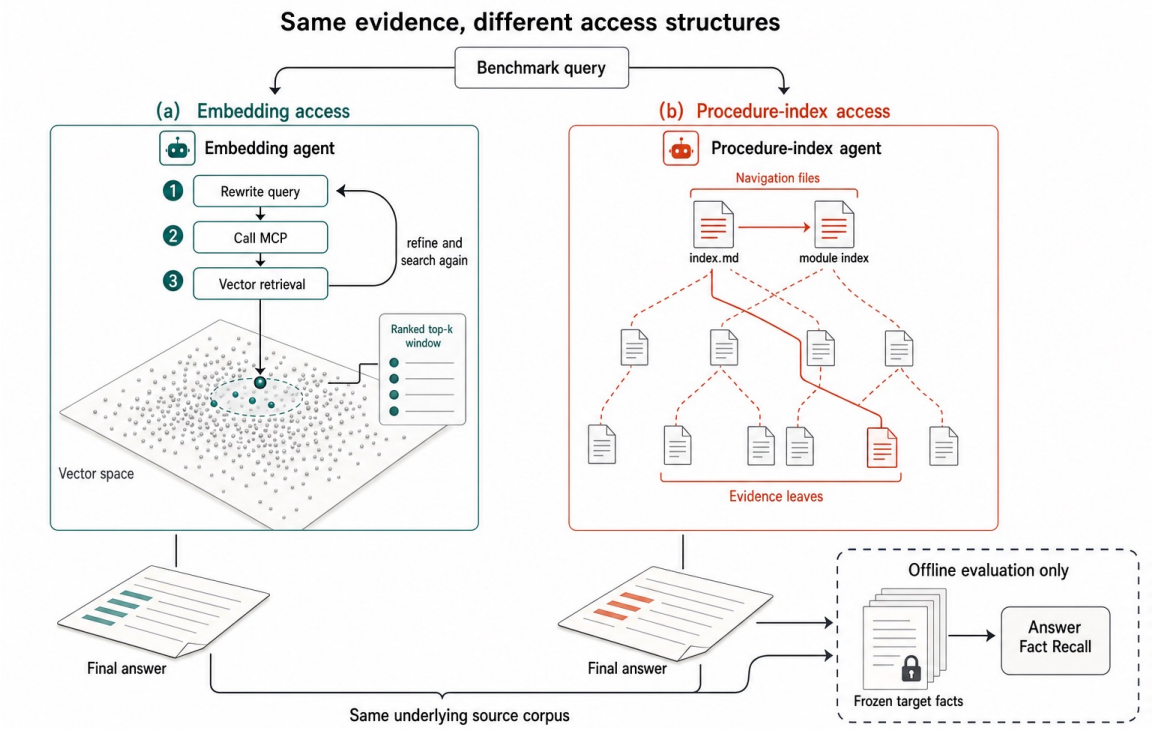

Figure 1: Same benchmark query and source corpus, different evidence-access structures. The embedding agent iteratively rewrites its query, calls an MCP vector-search tool, and receives a rank-truncated candidate window. The procedure-index agent reads navigation files and follows routes to source-linked evidence items. Both produce final answers evaluated offline against the same frozen target-fact set.

2 Related Work: From Retrieval to Navigable Evidence Access

2.1 Structured Retrieval Beyond Flat Chunks

A substantial line of work shows that flat similarity retrieval can be insufficient for corpus-level questions. Entity graphs and community summaries, graph-based non-parametric memory, and tree-organized hierarchical summaries all improve retrieval-augmented generation in their settings (Edge et al., 2024; Han et al., 2025; Jimenez Gutierrez et al., 2024; 2025; Sarthi et al., 2024; Zhao and Yang, 2026). Structure-aware retrieval over complex or book-length documents further shows the value of document hierarchy, and recent work shows that structure can also be constructed at inference time rather than fixed at indexing time (Wang et al., 2025; Huang et al., 2025; Li et al., 2025). This literature establishes corpus structure as an existing design dimension for evidence access. We study this design dimension through a procedure-structured interface for heterogeneous due diligence and evaluate it at the level of final-answer fact coverage.

2.2 Active Navigation and Agentic Retrieval

Classic IR and HCI models treat search as iterative navigation rather than a single lookup. Berrypicking, exploratory search, and information foraging describe evolving needs, informative cues, and movement through an information environment (Bates, 1989; Marchionini, 2006; Pirolli and Card, 1999). They provide a conceptual lens for our question: what evidence regions and next actions does an interface make observable at each step? Recent RAG systems show a parallel move from passive top-k lookup toward active navigation: agents that traverse knowledge structures, corpora compiled into agent-navigable directories and skill files, and hierarchical retrieval interfaces from which an agent chooses its own strategy (Dai et al., 2026; Sun et al., 2026; Du et al., 2026). A closely related recent system, Corpus2Skill, compiles

an enterprise corpus into a hierarchical skill directory, exposes that directory to an agent through progressive disclosure, and evaluates where navigation helps across enterprise QA collections (Sun et al., 2026). Procedure-index shares the compile-then-navigate pattern and filesystem-based progressive disclosure, but organizes evidence around a fixed analytical procedure and studies final-answer target-fact coverage, cumulative interface packages, and fixed-budget evidence access in heterogeneous due diligence.

NaviRAG studies active top-down navigation primarily within long documents (Dai et al., 2026). A-RAG exposes a hierarchical retrieval interface at corpus scale (Du et al., 2026). Psi-RAG builds a hierarchical abstract tree for cross-document retrieval (Zhao and Yang, 2026).

Navigation structure, prompt-level skill text, representation, and tool policy remain bundled in many agentic systems. Our sequential interface ablation therefore estimates adjacent package increments from hierarchical rendered evidence to navigation documents to skill guidance, without treating the design as a factorial isolation of every component.

2.3 Retrieval as Policy and Trajectory

Recent work treats retrieval as sequential decision making: multi-turn interaction with a graph environment optimized by reinforcement learning, structure-coherent intermediate retrieval paths with progress-aware rewards, and survey-level formulations of agentic retrieval as trajectory-level decision processes with cost and stopping decisions (Luo et al., 2025; Park et al., 2026; Singh et al., 2025; Mishra et al., 2026). Unlike policy-learning work, this study does not train or optimize retrieval control. It compares two agent systems that receive the same query and requested analytical coverage and use the same answer-model name; representation, tools, and interface-specific instructions remain part of each system. The study asks how a procedure-structured interface behaves in final-answer fact coverage relative to whole-document embedding search and uses selected traces to examine where that surface is or is not exploited.

2.4 Budget, Stopping, and Operating-Regime Evaluation

Evaluation of agentic retrieval requires more than answer accuracy because trajectories, budgets, and stopping behavior matter, and whether structure helps depends on the task and pipeline (Han et al., 2025; Xiang et al., 2025; Singh et al., 2025; Mishra et al., 2026). We therefore use fact-level answer coverage as the primary outcome, traces to diagnose where corpus evidence was or was not opened, a natural-stop budget sweep to examine sensitivity across operating points, and a forced-continuation follow-up to probe stopping behavior.

2.5 Fact-Level Answer Evaluation

Fact-level answer evaluation has two directions. Precision-oriented scoring decomposes the generated answer and verifies each claim against sources (Min et al., 2023; Wei et al., 2024). Recall-oriented scoring instead measures coverage of reference facts, a lineage running from nugget-based definition-question evaluation and pyramid summarization scoring to LLM-automated nugget recall in the TREC 2024 RAG track (Voorhees, 2003; Nenkova and Passonneau, 2004; Pradeep et al., 2024; 2025). The two directions capture complementary failure modes—unsupported content versus omission—and omission is the failure mode this study targets.

Existing recall-oriented protocols extract nuggets from pooled documents at evaluation time, which suits open web corpora with many participating systems, while closed-corpus QA benchmarks constrain questions to short answers where exact match suffices (e.g., Yang et al., 2018); the navigation systems above likewise evaluate at answer level. Neither protocol fits a paired comparison of two interfaces over a fixed collection with long-form, multi-fact answers: evaluation-time extraction ties the target set to the compared systems' output pools, and short-answer formats cannot express coverage. This setting motivates a frozen, pre-reviewed, query-scoped target set with priority tiers and a closed-world, span-quoting recognizer.

Table 1: Artifacts exposed by the procedure-index interface. "Carries evidence text" means evidence-bearing content, not benchmark targets.

Artifact	Form	Generation source	Carries evidence text	Agent action
Source record	Shared	Ingested registry or web document	Yes	Open the raw payload behind an evidence item
Evidence item	Shared	Rule-based rendering plus targeted extraction/routing	Yes	Open a concrete evidence item and follow its source pointer
Overview	Region	Template over module inventory and item metadata	Selected summaries	Scan a typed region before opening evidence items
Route document	Route	Deterministic hierarchy and event/entity grouping	Labels or short fields	Descend to a narrower evidence family
Evidence map	Region	Cross-module rule-based views over indexed items	Short evidence cues	Identify candidate regions across modules
High-signal view	Region	Heuristic selection from the evidence map	Short evidence cues	Inspect a compact first-read queue, then open evidence items

3 Procedure-index as a Hierarchical Navigation Interface

Procedure-index compiles a complex company document collection into a navigable evidence environment for an agent. Rather than exposing its internal directory schema, the interface orients the agent in three forms: analytical regions for broad task areas, cross-cutting entity routes for following companies and related actors, and chronological or event routes for tracking developments over time. Analytical regions are realized as a small fixed set of typed modules; later sections use "typed module" in this sense. Using read-only tools, the agent can move from this overview to narrower evidence views and then to source-linked evidence items. Table 1 lists the artifacts that realize these three forms and the agent action each supports.

The agent navigates this environment through three to four levels of read-only file access: from a small set of typed-module overviews, through region- or route-level navigation documents, down to individual evidence items linked to their source payloads. Typed modules number in the low tens and are shared across companies. Route documents—entity hierarchies and chronological groupings—number in the tens to low hundreds per company. Evidence items, produced by rule-based rendering of source payloads combined with targeted extraction and routing into the appropriate regions, reach the tens of thousands across the full corpus. The agent operates with file-read and directory-list tools only; it cannot write files, call a search API, or access the web. High-signal views apply heuristic filtering over the evidence map to surface a compact first-read queue; they are not learned rankings.

Each source-linked evidence item retains a provenance link to its underlying payload. The access layer changes both organization and visible content: evidence items carry evidence text, while navigation documents carry labels, short fields, or evidence cues. The sequential interface ablation therefore separates the evidence-bearing layer from the navigation layer without claiming that navigation is content-free. Section 6 examines a cue-bearing view, and Appendix A.7 reports the translated, partially redacted prompt contracts used by the two main arms.

Figure 1 contrasts the two system interfaces. The embedding agent rewrites its query and calls an MCP vector-search tool, which returns the top-k whole documents by cosine similarity. The procedure-index agent instead starts from navigation files and descends their routes to source-linked evidence items through local file tools. Both derive from the same underlying source environment, but they differ in representation, visible artifacts, tool protocol, and main-experiment budget. The sequential interface ablation and stopping-aware budget analyses examine these bundled differences below.

The region-and-route schema is fixed in code rather than generated from the evaluation queries or target fact set, and the same procedure-structured interface is reused across all query types. We treat the remaining build logic as part of the evaluated system implementation rather than as an experimental variable in this paper.

4 Experimental Setup

4.1 Benchmark

We evaluate agents on a due-diligence company fact benchmark with 50 cases: 10 companies crossed with five query types (investment/M&A due diligence, credit and supplier access, legal compliance, investigation tree, and industry-sales analysis). The investigation-tree query asks for an evidence-backed timeline of founders, core team, shareholders, control chain, related parties, key events, negative coverage, penalties, and litigation. The frozen target set contains 3,321 query-scoped facts: 584 critical, 1,719 important, and 1,018 relevant. Each fact is an answer target for one company-query pair, not a document-hit target. Critical means that omission would materially weaken the query's due-diligence conclusion; important supports a key part of the answer; relevant supplies lower-priority context. Appendix A documents benchmark construction, review, and the target-set provenance boundary.

The realized snapshot, shared by both arms, contains 29,688 Chinese-language documents for 10 companies, assembled from access-controlled, internally maintained sources, including licensed corporate-registry records. These statistics describe the frozen materialized snapshot rather than an upstream collection-success rate: this release does not include a complete record of attempted, failed, or rejected collection operations. Appendix A.8 reports per-company statistics. We study this local collection because due-diligence answers must be checked against a fixed evidence base; public-web results drift, and relevant registry material is licensed. The median document contains 556 characters, and the 90th-percentile length is 2,717 characters. Agents receive the query and document environment but do not receive the target facts.

4.2 Systems Compared

Both main arms use the same endpoint model identifier, `deepseek-v4-flash`, through the Claude Code CLI harness and a DeepSeek Anthropic-compatible endpoint.

The embedding arm is an agentic operational comparator: the answer agent iteratively searches and fetches from a flat whole-document vector store through a Model Context Protocol tool interface. Whole-document indexing is a practical retrieval unit at the document lengths reported in Section 4.1. `BAAI/bge-m3` embeddings (Chen et al., 2024) are 1024-dimensional, one vector per whole document with no sub-chunking and an 8,000-character indexing cap, searched by exact cosine similarity. The runtime contract permits at most 6 *successful* search calls, default top-$k = 10$ with a per-call maximum of 20, at most 30 fetched documents, and 4,000 characters per fetch.

The procedure-index interface exposes the full navigation surface of Section 3, plus skill guidance, through read-only local tools. It has no per-call retrieval cap and stops when it judges the task complete, subject to a one-hour timeout that no published run reaches. This main-protocol asymmetry is part of the system-level treatment.

The budget analysis adds three protocol layers. First, the initial pilot comparison applies a shared $0.40 evidence-collection cap to both complete interfaces and then performs tool-free synthesis. Second, the natural-stop sweep applies nominal evidence-collection caps from $0.40 to $2.00 while preserving each arm's ordinary stopping behavior; the embedding sweep removes fixed search-call and fetch-document limits, so the nominal cap is the binding resource control. Third, an embedding-only follow-up forces evidence collection to continue across ordinary end turns until the budget terminal, after which the same tool-free synthesis runs once. The full protocol and budget grids appear in Appendix A.12.

4.3 Scope of Claims

The primary contrast is procedure-index versus embedding under the same query set, requested analytical coverage, frozen target set, underlying source environment, answer-model name, and answer-fact recognizer. Decoding parameters were left at endpoint defaults and were not pinned, and the two arms' answers come from separate run generations—answer batches executed at different times—rather than one concurrent randomized batch. Everything else varies with the arm—representation, tool protocol, interface-specific instructions, access budget, stopping policy, and run generation—so we interpret the

paired contrast as the observed difference between the two deployed interfaces as complete systems, not as an isolated effect of index structure. The sequential interface ablation probes cumulative interface packages. The natural-stop analysis remains a complete-interface comparison across nominal caps, and the forced embedding follow-up changes both stopping instructions and run generation; neither isolates every factor.

The main procedure-index row also has a separate provenance limitation. It is rescored from an imported historical answer batch, and run identifiers recorded as sources for some admitted benchmark candidates also occur in that historical batch. Because the benchmark metadata did not retain answer digests, this overlap establishes provenance overlap, not byte-identical answers or same-run target exposure. We therefore do not treat the main row as an independent validation of benchmark construction (Appendix A).

This scope governs all results below. Each case runs once per arm at each reported point, so intervals and paired tests reflect variation across the 10 company clusters rather than within-arm generation variance. All 50 cases are present for each primary arm. The primary tables use scored-fact denominators; Appendix A.5 reports the fact-level unscored counts and conservatively counts every unscored decision as a miss over all 3,321 targets.

4.4 Metrics and Scoring

The primary outcome is answer fact recall. Strict recall gives credit only when the final answer fully expresses a target and its indispensable anchors. For example, if a target states that a company received a regulatory penalty in 2024, strict credit requires the answer to preserve the company, event, and year. Near-hit recall also credits a non-contradicted partial expression, such as mentioning the penalty while omitting a required anchor. An incompatible year, entity, or direction is a contradiction and receives no near-hit credit; a decision is unscored when the recognizer cannot apply the contract reliably. Let D_S be the scored target facts in scope S, where S is either the critical subset or all priority levels. Let F_S count fully expressed targets and P_S count partially expressed, non-contradicted targets. Then

$$R_{\text{strict}}(S) = \frac{F_S}{|D_S|}, \qquad R_{\text{NH}}(S) = \frac{F_S + P_S}{|D_S|}.$$

The primary denominator excludes decisions marked unscored; Appendix A.5 instead uses the complete frozen target scope and assigns unscored decisions zero credit. Both views score expression in the final answer text against the frozen targets: they are agnostic to which documents were retrieved and require no citation support.

Final answers are scored by the same closed-world recognizer, `deepseek-v4-pro`. It evaluates only the predefined fact keys and first records whether a decision is scored or unscored. Each scored decision contains a three-way coverage label—full, partial, or none—and an independent contradiction flag; partial coverage may coexist with a contradiction, while full coverage may not. For full or partial coverage, or for a positive contradiction flag, the recognizer must quote a verbatim span from the answer. It cannot add targets or issue a holistic answer preference. Relative to an open-ended LLM-as-judge setup, this reduces freedom to invent criteria or redefine relevance and makes positive and contradictory decisions traceable to predefined targets and answer spans (Zheng et al., 2023; Wang et al., 2024). It does not eliminate model bias or prompt sensitivity. Re-scoring every main-comparison answer with both the identical recognizer configuration and an independent GLM-5.2 judge leaves all four planned contrast directions and significance conclusions unchanged (Appendix A.10).

4.5 Statistical Tests

The 50 company-query cases are query observations nested in 10 company clusters, each containing the same five query types. Point estimates pool eligible scored facts. We compute 95% percentile confidence intervals from 10,000 cluster-bootstrap resamples, drawing 10 companies with replacement, carrying all five within-company queries with each draw, and recomputing pooled fact recall; there is no separate query-type stratification. The bootstrap seed is fixed at 20260709 (Efron and Tibshirani, 1993). Ablation adjacent-increment intervals use the same company-cluster design with a stable SHA-256-derived seed for each arm pair and metric.

Table 2: Main comparison: procedure-index records higher recall on all four primary metrics. Cells report pooled fact recall with 95% company-cluster bootstrap CIs.

Metric	Embedding	Procedure-index	Difference	p
Critical near-hit	0.400 [0.356, 0.438]	0.550 [0.492, 0.628]	**+0.150**	**0.006**
Critical strict	0.130 [0.096, 0.166]	0.269 [0.229, 0.317]	**+0.139**	**0.004**
Overall near-hit	0.270 [0.235, 0.301]	0.405 [0.366, 0.444]	**+0.134**	**0.002**
Overall strict	0.083 [0.070, 0.097]	0.195 [0.172, 0.221]	**+0.112**	**0.002**

Note: Both arms contain 50 runs; difference is Procedure-index minus Embedding; recall denominators include scored facts only; planned contrasts use exact two-sided company-cluster permutation p-values; p-values are nominal and unadjusted; bold marks nominal $p < 0.05$; legacy public-web reference rows appear in Appendix Table 6.

Pairwise contrasts use the pooled fact-recall difference shown by the corresponding point estimates and a two-sided company-level label-permutation test (Good, 2000). One label flip moves all five queries for a company together; with 10 companies, we enumerate all 2^{10} assignments and recompute both arms' pooled numerators and denominators under each assignment. Ablation ordered trends use 10,000 label permutations within each company, the predeclared Flat (000) < Hierarchical evidence (100) < Navigation documents (110) < Skill guidance (111) order, pooled arm rates, a plus-one correction, and a one-sided greater test. All planned p-values are nominal and unadjusted.

5 Results

5.1 Main Comparison: Final-Answer Fact Coverage

Coverage. Table 2 shows the central result. Relative to the embedding interface, procedure-index records critical near-hit recall of 0.550 versus 0.400 and overall near-hit recall of 0.405 versus 0.270; critical strict is 0.269 versus 0.130 and overall strict is 0.195 versus 0.083. The four exact company-cluster tests range from $p = 0.00195$ to $p = 0.00586$. Counting every unscored decision as a miss reduces the estimates but preserves all four gaps at +0.106 to +0.124 (Appendix A.5). On this benchmark, the complete procedure-index interface covers more target facts than the embedding baseline.

Table 2 also carries two historical reference rows whose discovery surface includes the public web. They are stronger on some metrics, especially critical near-hit recall; they measure what public-web discovery can reach over this benchmark and appear as reference context in Appendix A.9. Their different discovery surface also illustrates why they cannot substitute for the fixed, access-controlled local setting studied here.

Recorded cost. The main protocol reports an effectiveness-cost pair, not a cost-normalized advantage. Mean recorded usage is about 1.26M model-API tokens per procedure-index run versus about 198K per embedding run, a factor of about 6.3; the totals sum input, cache-read, cache-creation, and output tokens across each run's model calls. Procedure-index runs average about 150 seconds of wall-clock time across all 50 runs; the embedding runner recorded no comparable timings. Case-level model-API spending has the same direction (Section 6). These figures exclude one-time construction: recorded index builds took roughly five to six minutes per company, amortized across five query types, while comparable embedding build accounting is unavailable. We therefore do not impute either unobserved build cost. The budget analyses revisit the contrast across nominal evidence-collection caps and stopping policies.

5.2 Sequential Interface Ablation

The sequential interface ablation is an ordered, cumulative design over three agent-facing packages: hierarchical rendered evidence (H), cue-bearing navigation documents (N), and skill guidance (S). The observed profiles are *Flat (000)* (historical `base`), *+Hierarchical evidence (100)* (`no-navigation`), *+Navigation*

Table 3: Sequential interface ablation: adjacent H/N/S package increments. Difference cells report pooled fact-recall increments with 95% company-cluster bootstrap CIs.

Package added	Metric	Reference	With package	Difference [95% CI]	p
Hierarchical evidence	Critical near-hit	0.382	0.418	+0.036 [-0.021, +0.082]	0.256
	Critical strict	0.125	0.180	**+0.055** [+0.020, +0.090]	**0.021**
	Overall near-hit	0.290	0.336	+0.046 [+0.008, +0.087]	0.064
	Overall strict	0.100	0.136	**+0.037** [+0.014, +0.060]	**0.014**
Navigation documents	Critical near-hit	0.418	0.551	**+0.133** [+0.084, +0.185]	**0.004**
	Critical strict	0.180	0.217	+0.037 [-0.023, +0.099]	0.389
	Overall near-hit	0.336	0.399	**+0.063** [+0.026, +0.102]	**0.016**
	Overall strict	0.136	0.148	+0.012 [-0.017, +0.039]	0.453
Skill guidance	Critical near-hit	0.551	0.558	+0.007 [-0.019, +0.033]	0.623
	Critical strict	0.217	0.240	+0.023 [-0.017, +0.068]	0.318
	Overall near-hit	0.399	0.409	+0.010 [-0.013, +0.035]	0.463
	Overall strict	0.148	0.161	+0.013 [-0.002, +0.029]	0.168

Note: Each adjacent comparison contains 50 paired cases; H compares profile 100 with 000, N compares 110 with 100, and S compares 111 with 110. H bundles hierarchy with rendered leaves, N bundles route organization with navigation cues, and S bundles skill guidance with its skill-aware prompt. Difference is with-package minus reference; recall denominators include scored facts only; intervals resample 10 company clusters with all five within-company queries retained; exact two-sided p-values use all 2^{10} company-level label permutations and are nominal and unadjusted; bold marks nominal $p < 0.05$. Non-significant increments do not establish equivalence.

documents (110) (`no-skill`), and +*Skill guidance (111)* (`index-agent-full`). Their adjacent estimands are $\Delta_H = Y_{100} - Y_{000}$, $\Delta_{N|H} = Y_{110} - Y_{100}$, and $\Delta_{S|H,N} = Y_{111} - Y_{110}$.

These are conditional package increments, not factorial main effects or interactions: H adds both source hierarchy and rendered evidence items, N adds both route organization and short navigation cues, and S adds both skill guidance and its skill-aware prompt. All four profiles share the same 50 cases; the three non-full profiles use the same neutral prompt.

Figure 2 shows the ordered estimates; Appendix Table 8 reports every profile and interval. Profile 111 is an independent rerun of the main procedure-index environment from a later run generation; its critical near-hit and strict estimates are 0.558 and 0.240 versus 0.550 and 0.269 in the main run, which we read as an informal stability check only.

Hierarchical-evidence increment. Table 3 reports all three adjacent contrasts. Adding hierarchical evidence (100 minus 000) increases critical strict recall by +0.055 (95% CI [+0.020, +0.090], $p = 0.021$) and overall strict recall by +0.037 ([+0.014, +0.060], $p = 0.014$). The critical and overall near-hit increments are +0.036 ($p = 0.256$) and +0.046 ($p = 0.064$), respectively, and are inconclusive under the nominal exact tests.

Navigation-document increment. Adding the navigation-document package (110 minus 100) increases critical near-hit by +0.133 (95% CI [+0.084, +0.185], $p = 0.004$) and overall near-hit by +0.063 ([+0.026, +0.102], $p = 0.016$). The strict increments are +0.037 and +0.012, with intervals crossing zero and $p = 0.389$ and $p = 0.453$. The package therefore has an observed adjacent increment for near-hit recall, while the strict increments remain inconclusive.

Skill-guidance increment. The 111-minus-110 contrast estimates the skill-guidance package increment; its non-significant tests do not establish equivalence. On the 3,321-fact benchmark, the predeclared ordinal trend tests are below 0.001 for all four metrics. These tests pool evidence across the whole sequence, so they can be significant even where a single adjacent step is not.

In this cumulative sequence, the observed strict-recall increase occurs when the hierarchical-evidence package is added to flat raw evidence, while the observed near-hit increase occurs when the navigation-document package is added to hierarchical rendered evidence. Because profile 100 already holds every

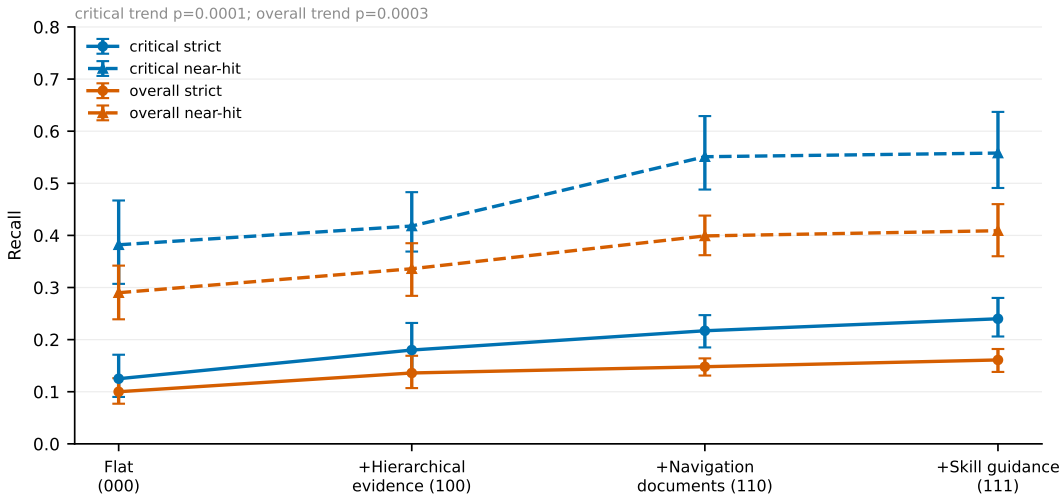

Figure 2: Sequential interface ablation. The largest observed adjacent near-hit increase occurs when the navigation-document package—route organization plus short evidence cues—is added; points show pooled recall estimates with 95% company-cluster bootstrap CIs.

rendered evidence item, the navigation increment is not explained by evidence re-rendering alone; it estimates the joint contribution of route organization and the short evidence cues those documents carry. The data do not show that skill guidance is ineffective or equivalent to profile 110. This pattern is consistent with pipeline evidence that hierarchy and traversal work jointly (Dai et al., 2026), within the limits of our adjacent contrasts.

5.3 Budget Sensitivity and Stopping

Initial $0.40 pilot operating point. The original paired check asks whether the procedure-index signal remains visible when both interfaces use a shared $0.40 evidence-collection cap. Overall near-hit recall is 0.297 for procedure-index and 0.254 for embedding; the planned exact company-cluster test is positive but suggestive ($p = 0.059$). Overall strict is 0.097 versus 0.084 ($p = 0.324$), and the two critical-recall contrasts are not significant. This anchor is a separate run generation from the multi-budget analyses below; Appendix Table 9 reports all four estimates and planned tests.

Natural-stop budget sensitivity. Figure 3 reports the 10 shared nominal caps from $0.40 through $2.00. Procedure-index is descriptively higher than natural-stop embedding at every plotted cap on both metrics. At $0.40, strict recall is 0.119 versus 0.077 (gap +0.042) and near-hit recall is 0.385 versus 0.284 (+0.101). At $2.00, strict recall is 0.206 versus 0.097 (+0.109) and near-hit recall is 0.446 versus 0.309 (+0.137). These are pooled scored-fact estimates, not pointwise significance claims or equal-realized-cost comparisons.

The nominal cap changes whether the natural-stop embedding runs encounter the budget terminal, but it does not prescribe acquisition actions. All 50 embedding runs stop with `budget_exhausted` at $0.40; none do so from $1.30 through $2.00. Aggregate search calls change from 444 at $0.40 to 476 at $2.00, and fetched documents from 1,017 to 1,290. Thus, above the point where the cap ceases to bind, raising the nominal cap alone does not force additional acquisition.

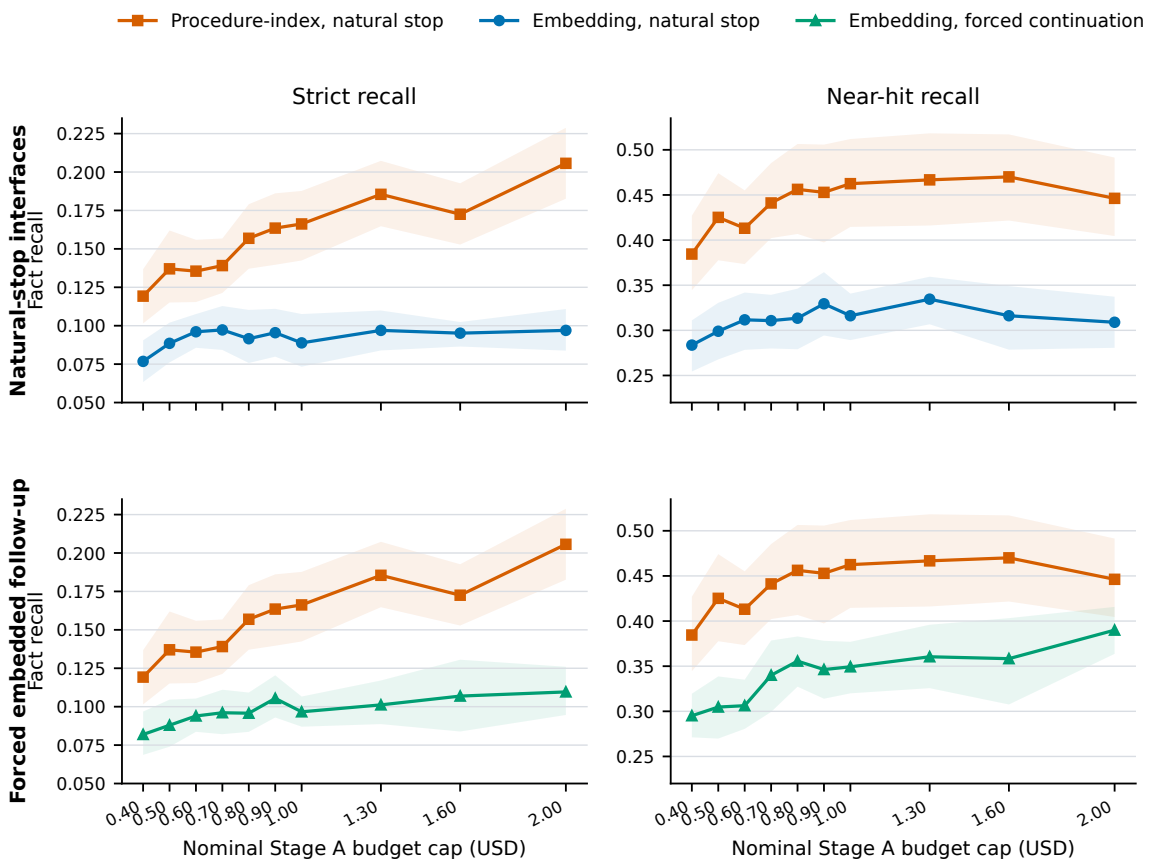

Figure 3: Budget sensitivity and stopping. The top row compares the two natural-stop interfaces at 10 shared nominal evidence-collection caps. The bottom row compares forced-continuation embedding with the same natural-stop procedure-index line as a descriptive reference. Points are pooled scored-fact recall over 3,321 targets; shaded bands are 95% company-cluster bootstrap CIs over 10 companies. Each case is generated once per point.

Forced-continuation follow-up. The embedding-only follow-up changes that stopping behavior: after an ordinary end turn, evidence collection must continue and fetch new evidence until reaching the budget terminal. Across the $0.40 and $2.00 endpoints, mean realized evidence-collection spending rises from $0.440 to $2.046, mean search calls from 9.20 to 23.52, and mean fetch calls from 2.90 to 10.00. This verifies separation in realized acquisition effort, not equal evidence quantity across interfaces.

Across the 11-point forced-continuation grid, the observed pooled recall estimates rise overall but are not monotonic. With one generation at each point, this grid does not establish a stable non-monotonic budget-response curve. From $0.40 to $2.00, strict recall changes from 0.082 to 0.110 (+0.028) and near-hit recall from 0.295 to 0.390 (+0.095). At $2.00, the repeated natural-stop procedure-index reference is 0.206 strict and 0.446 near-hit, leaving descriptive gaps of +0.096 and +0.056. Because there is no forced procedure-index arm and the protocols differ, these gaps do not estimate a structure-only effect or a matched cost-efficiency contrast.

6 Qualitative Analysis

6.1 Case Selection

The aggregate results show a fact-coverage difference; they do not identify its mechanism or establish that recalled facts were supported by opened evidence. This section asks whether available traces are

consistent with one qualitative account and shows its boundary. The cases were selected post hoc—Case A as a traceable success regime, Case C as the largest fixed-budget per-question gap, and Case D as the largest main-comparison reversal—to illustrate and bound the account, not to test it. Scores come from run artifacts; case names are pseudonymized and routes are coarsened. A fourth candidate was dropped after recognizer validation changed seven of its fact decisions; the remaining exhibits keep their original Case A, C, and D codenames. We report the selection rules, metrics, aggregate access observations, alternatives checked, and verification boundary, but do not release path-level case artifacts.

6.2 Selected Dispersed-Evidence Cases

Each embedding step writes a semantic query, receives up to the requested top-k documents, fetches selected results, and repeats; at every step, the visible surface is the returned rows. Each index step reads a directory or navigation document, descends into a typed module, and opens evidence items; navigation artifacts expose more of the corpus organization before item selection.

Top-k retrieval requires each document to rank highly before the agent can see it. This can favor overview material that resembles broad company-level queries across repeated searches. Decision-relevant facts often appear instead in event-specific documents, each matching only a narrow part of the query and remaining below the visible cutoff. Structured traversal reduces this dependence on rank competition for typed evidence regions. A single financing event, for example, need not outrank overview documents if traversal reaches a financing-history route; visibility then depends partly on collection structure rather than only on similarity score.

The interfaces also expose different information about what has not been inspected. The embedding agent sees only returned rows. The index agent can inspect module existence and route inventories, although an unvisited or weakly aligned evidence item still prevents a source-grounded conclusion.

A dispersed event-level case (Case A). Critical near-hit is 0.333 for embedding and 0.750 for procedure-index. Documents associated with the index answer's distinctive facts exist in the shared frozen snapshot and in the embedding vector store, yet appear in none of the embedding run's search observations or fetches. That run completed six successful broad searches centered on company overview, financial performance, and market position. The procedure-index run instead visited ownership, financing, risk, and cross-module evidence-map regions, then covered financing and regulatory facts missing from the embedding answer. At this case level, the observed gap occurs before document selection and is consistent with structured navigation making a narrow evidence region visible.

A fixed-budget dispersed-evidence case (Case C). Critical near-hit is 0.273 for embedding and 0.909 for procedure-index. The embedding-visible material emphasized neutral company descriptions, while negative turning-point facts appeared only after index traversal. The embedding answer produced a uniformly positive profile and reported sales-relevant leads as "not covered." The index trace's high-signal view—itself a cue-bearing artifact, not just a path—exposed a failed public test, a layoff plan, and an industry-ranking drop, which its answer turned into sales entry points. This is the largest per-question gap at the shared $0.40 cap; as a selected extreme, it shows the same trace pattern can occur even at the lowest cap in the grid.

6.3 A Ranking-Favorable Boundary Case

The aggregate advantage is not uniform across questions. In critical near-hit, procedure-index is higher on 39 of 50 main cases, tied on 2, and lower on 9; only 3 of 50 also reverse on overall near-hit. Under the fixed-budget evaluation, the critical near-hit reversal count rises to 19 of 50 while the pooled difference remains positive. These exploratory counts motivate inspection of where the qualitative account fails.

The largest full reversal is Case D, industry-sales analysis: critical near-hit is 0.526 for embedding versus 0.238 for procedure-index, with zero contradicted facts on both sides. The reversal survives counting the embedding agent's two unscored critical facts as misses (0.476 vs. 0.238). Here, several target facts concentrate in a few query-similar public articles; the most fact-dense annotation-side source supports

4 of the question's 21 critical facts. The embedding agent reached this material in one fetch sequence. Procedure-index navigation instead landed on coarser partner and customer-record regions. When a few query-similar articles already concentrate the answer, ranking concentration can be a shortcut and route coverage a detour.

The reversal is consistent with the same account in the opposite regime: typed modules offer less advantage when they contain coarse records and the needed facts are concentrated in a few query-similar articles. This selected case motivates, but does not test, the hypothesis that route granularity matters differently across tasks.

6.4 Verification Boundary

For these qualitative cases we verified corpus parity (the associated documents exist in the embedding vector store and match document ids) and model-name parity, confirmed the absence of web search in the index traces, and checked truncation (the three documents inspected are 184–3,740 characters, below the 4,000-character fetch cap; 6.4% of corpus documents exceed that cap). The main budget asymmetry remains real, roughly $0.43 for the embedding run versus $1.67 for the procedure-index run in Case A, which is why no case study is used to convert the main system-level association into a structure-only causal effect.

7 Discussion

7.1 When Structured Navigation Helps

Across these qualitative cases, the recurring pattern is dispersion. In the selected cases, procedure-index's higher near-hit recall coincides with target facts being spread across typed regions while broad similarity searches return overview material. Case A illustrates this event-level setting; Case C shows that the same pattern can occur under the fixed-budget constraint. The navigation surface contributes orientation and short cues, while the agent still decides which evidence items to open and which facts to retain. This division of labor is consistent with route organization helping without implying that every navigation policy is effective.

7.2 Scope and Operating Regimes

Case D marks the opposite regime: when a few query-similar articles concentrate several target facts, direct ranking can reach the useful material faster than a coarse route. It suggests a practical boundary for future tests: the value of structured navigation may depend on whether its route granularity matches the task's fact distribution. A stronger chunked, hybrid, or reranked baseline may shift that boundary.

7.3 Policy Under Budget

The budget experiments expose two policy questions central to agent-visible evidence access: how an interface allocates a nominal spending cap across actions of different granularity, and when it decides that further acquisition is unnecessary. Forcing the embedding agent to continue separates acquisition effort from stopping behavior: realized spending and search/fetch activity rise while recall does not follow a stable curve (Section 5.3). The nominal cap and the stopping rule are therefore distinct parts of the evaluated interface, and an interface that stops early leaves a higher cap unused.

Future policies could allocate finite spending across high-value unvisited routes, detect when a fact-dense ranked article makes further traversal unnecessary, and reserve enough context for synthesis. Evaluation should pair those policies with fact-to-opened-source attribution so changes in answer coverage can be distinguished from unsupported recall and from repeated access to redundant evidence.

8 Limitations and Open Questions

Baseline. The baseline is conservative: whole-document BGE-M3 without chunking, hybrid retrieval, or reranking, under a capped search protocol in the main comparison. A stronger retrieval stack could shift the observed gap. The natural-stop sweep compares complete interfaces rather than equal realized spending. The forced-continuation sweep is embedding-only, reuses the natural-stop procedure-index curve as a descriptive reference, and changes stopping instructions together with run generation; it therefore cannot attribute the remaining gap to structure. The study does not directly compare procedure-index with a Corpus2Skill-like generic navigation hierarchy and therefore does not identify procedure alignment itself as the cause of the observed system-level difference.

Generation and inference. Each case ran once per arm at each reported point with unpinned endpoint-default decoding. Company-cluster inference preserves dependence among the five queries for each company, but with only 10 clusters it is coarse and does not measure generation variance; generation noise could move individual estimates in either direction. We do not conduct pointwise tests across the sweep. The natural-stop grid was adaptively enriched after review of a coarser completed grid, and the forced grid adds a $1.20 point not shared by the natural-stop curves.

Benchmark and measurement. The main procedure-index row is rescored from a historical answer batch with run-identifier overlap in benchmark provenance and is not an independent validation of benchmark construction (Section 4.3). Critical-target priorities received a complete human quality pass, but the review records do not support independent duplicate annotation, formal adjudication, or inter-annotator agreement. The LLM recognizer has same-configuration and independent-judge retests but no sampled human audit. Fact recall is neither precision nor citation-supported recall, so parametric memory may contribute.

Scope and reproducibility. The main treatment is system-level, and the benchmark covers one Chinese-language company due-diligence domain whose licensed fact statements cannot be released. Nominal evidence-collection cap, realized model spending, search or fetch actions, and evidence quantity are different quantities; the present data do not support a common cost–recall frontier. Case and subgroup analyses are post hoc, and cost accounting omits comparable build cost. Generalization, stronger retrieval baselines, navigation-design search, and fact-to-source attribution remain future work.

9 Conclusion

In this access-controlled company due-diligence setting, the complete procedure-index interface records higher final-answer fact coverage than the whole-document embedding baseline across all four primary metrics, at substantially higher recorded model usage. The sequential interface ablation locates the strict-recall increase at the hierarchical-evidence step and the largest adjacent near-hit increase at the navigation-document step. Across 10 shared natural-stop budget caps, procedure-index remains descriptively higher on both metrics; forcing the embedding agent to continue acquisition does not close the gap. Selected cases place the advantage in a dispersed-evidence regime and its ranking-favorable boundary. Scope, limitations, and open questions are discussed in Sections 4.3 and 8.

Availability. This preprint release contains the paper source, generated tables, bibliography inputs, and rendered figures. It does not contain the supporting experimental code, internal Answer skill, or traversal sequence, so it supports interpretation of the reported aggregates but not complete reproduction of the procedure-index arm. The licensed document collection and benchmark fact statements cannot be released; the appendix documents their construction and statistics. Raw answers, raw judge output, raw traces, full case files, local paths, licensed documents, and benchmark fact text are not part of this release.

References

Marcia J. Bates. 1989. The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review* 13, 5 (1989), 407–424. https://doi.org/10.1108/eb024320

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, 2318–2335. https://doi.org/10.18653/v1/2024.findings-acl.137

Jihao Dai, Dingjun Wu, Yuxuan Chen, Zheni Zeng, Yukun Yan, Zhenghao Liu, and Maosong Sun. 2026. NaviRAG: Towards Active Knowledge Navigation for Retrieval-Augmented Generation. https://doi.org/10.48550/arXiv.2604.12766 arXiv:2604.12766 [cs.CL]

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. https://doi.org/10.48550/arXiv.2602.03442 arXiv:2602.03442 [cs.CL]

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. https://doi.org/10.48550/arXiv.2404.16130 arXiv:2404.16130 [cs.CL]

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

Phillip I. Good. 2000. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses* (2 ed.). Springer.

Haoyu Han, Li Ma, Yu Wang, Harry Shomer, Yongjia Lei, Zhisheng Qi, Kai Guo, Zhigang Hua, Bo Long, Hui Liu, Charu C. Aggarwal, and Jiliang Tang. 2025. RAG vs. GraphRAG: A Systematic Evaluation and Key Insights. https://doi.org/10.48550/arXiv.2502.11371 arXiv:2502.11371 [cs.IR]

Haoyu Huang, Yongfeng Huang, Junjie Yang, Zhenyu Pan, Yongqiang Chen, Kaili Ma, Hongzhi Chen, and James Cheng. 2025. Retrieval-Augmented Generation with Hierarchical Knowledge. https://doi.org/10.48550/arXiv.2503.10150 arXiv:2503.10150 [cs.CL]

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. In *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.2405.14831 arXiv:2405.14831 [cs.CL]

Bernal Jimenez Gutierrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. From RAG to Memory: Non-Parametric Continual Learning for Large Language Models. In *Proceedings of the International Conference on Machine Learning*. https://doi.org/10.48550/arXiv.2502.14802 arXiv:2502.14802 [cs.CL]

Zhuoqun Li, Xuanang Chen, Haiyang Yu, Hongyu Lin, Yaojie Lu, Qiaoyu Tang, Fei Huang, Xianpei Han, Le Sun, and Yongbin Li. 2025. StructRAG: Boosting Knowledge Intensive Reasoning of LLMs via Inference-Time Hybrid Information Structurization. In *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2410.08815 arXiv:2410.08815 [cs.CL]

Haoran Luo, Haihong E, Guanting Chen, Qika Lin, Yikai Guo, Fangzhi Xu, Zemin Kuang, Meina Song, Xiaobao Wu, Yifan Zhu, and Luu Anh Tuan. 2025. Graph-R1: Towards Agentic GraphRAG Framework via End-to-End Reinforcement Learning. https://doi.org/10.48550/arXiv.2507.21892 arXiv:2507.21892 [cs.CL]

Gary Marchionini. 2006. Exploratory Search: From Finding to Understanding. *Commun. ACM* 49, 4 (2006), 41–46. https://doi.org/10.1145/1121949.1121979

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.48550/arXiv.2305.14251 arXiv:2305.14251 [cs.CL]

Saroj Mishra, Suman Niroula, Umesh Yadav, Dilip Thakur, Srijan Gyawali, and Shiva Gaire. 2026. SoK: Agentic Retrieval-Augmented Generation (RAG): Taxonomy, Architectures, Evaluation, and Research Directions. https://doi.org/10.48550/arXiv.2603.07379 arXiv:2603.07379 [cs.AI]

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. 145–152. https://aclanthology.org/N04-1019/

Jinyoung Park, Sanghyeok Lee, Omar Zia Khan, Hyunwoo J. Kim, and Joo-Kyung Kim. 2026. HyperGraphPro: Progress-Aware Reinforcement Learning for Structure-Guided Hypergraph RAG. https://doi.org/10.48550/arXiv.2601.17755 arXiv:2601.17755 [cs.CL]

Peter Pirolli and Stuart Card. 1999. Information Foraging. *Psychological Review* 106, 4 (1999), 643–675. https://doi.org/10.1037/0033-295X.106.4.643

Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. 2024. Initial Nugget Evaluation Results for the TREC 2024 RAG Track with the AutoNuggetizer Framework. https://doi.org/10.48550/arXiv.2411.09607 arXiv:2411.09607 [cs.IR]

Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. 2025. The Great Nugget Recall: Automating Fact Extraction and RAG Evaluation with Large Language Models. https://doi.org/10.48550/arXiv.2504.15068 arXiv:2504.15068 [cs.IR]

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. In *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2401.18059 arXiv:2401.18059 [cs.CL]

Aditi Singh, Abul Ehtesham, Saket Kumar, Tala Talaei Khoei, and Athanasios V. Vasilakos. 2025. Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG. https://doi.org/10.48550/arXiv.2501.09136 arXiv:2501.09136 [cs.AI]

Yiqun Sun, Pengfei Wei, and Lawrence B. Hsieh. 2026. Don't Retrieve, Navigate: Distilling Enterprise Knowledge into Navigable Agent Skills for QA and RAG. https://doi.org/10.48550/arXiv.2604.14572 arXiv:2604.14572 [cs.IR]

Ellen M. Voorhees. 2003. Evaluating Answers to Definition Questions. In *Companion Volume of the Proceedings of HLT-NAACL 2003*. 109–111. https://aclanthology.org/N03-2037/

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 9440–9450. https://doi.org/10.18653/v1/2024.acl-long.511 arXiv:2305.17926 [cs.CL]

Shu Wang, Yingli Zhou, and Yixiang Fang. 2025. BookRAG: A Hierarchical Structure-Aware Index-Based Approach for Retrieval-Augmented Generation on Complex Documents. https://doi.org/10.48550/arXiv.2512.03413 arXiv:2512.03413 [cs.IR]

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form Factuality in Large Language Models. https://doi.org/10.48550/arXiv.2403.18802 arXiv:2403.18802 [cs.CL]

Zhishang Xiang, Chuanjie Wu, Qinggang Zhang, Shengyuan Chen, Zijin Hong, Xiao Huang, and Jinsong Su. 2025. When to use Graphs in RAG: A Comprehensive Analysis for Graph Retrieval-Augmented Generation. https://doi.org/10.48550/arXiv.2506.05690 arXiv:2506.05690 [cs.CL]

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/D18-1259 arXiv:1809.09600 [cs.CL]

Ziwen Zhao and Menglin Yang. 2026. Hierarchical Abstract Tree for Cross-Document Retrieval-Augmented Generation. In *Proceedings of the International Conference on Machine Learning*. https://doi.org/10.48550/arXiv.2605.00529 arXiv:2605.00529 [cs.LG]

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*. https://doi.org/10.48550/arXiv.2306.05685 arXiv:2306.05685 [cs.CL]

A Supplementary Benchmark Notes

A.1 Versioned-Liquid Benchmark and Frozen v2 Timeline

The benchmark follows a versioned-liquid lifecycle. Between releases, new evidence and query-relevant facts may update the evaluation fact set. Within a formal evaluation, however, the target snapshot is fixed and is not exposed to the agents. The present study uses v2 as its sole target snapshot. V2 was versioned and frozen on June 23, 2026. It contains 3,321 query-scoped answer targets over 10 companies, five query types, and 50 cases: 584 critical, 1,719 important, and 1,018 relevant. All reported recall estimates draw from this fixed v2 target set. For each arm, primary estimates use that arm's scored target facts in the relevant scope as the denominator. The main-comparison sensitivity analysis instead assigns zero credit to unscored decisions over all frozen targets in the corresponding scope.

Construction proceeded through versioned stages. The initial stage assembled 2,665 targets in total from the pre-collected evidence base, constructed separately for each of the 50 company-query records. The schema required every candidate to contain one atomic answer statement, detection words, at least one evidence record, its query scope, a priority label, and a merge key. The repository retained the final records, schema, manifest, and review notes, but intentionally excluded the original generation prompts, model streams, and runner scripts. We can therefore audit the resulting fields and evidence records, but cannot fully reconstruct the first-pass candidate-generation procedure or model from the public versioned artifacts.

The available construction records characterize this seed as primarily web-search-led. The evaluated collection, by contrast, integrates multiple non-public, internally maintained data sources. Closed-world scoring gives no credit to a valid, query-relevant fact that is absent from the current target set. The enrichment path addresses such reference-set omissions in later releases. Answer text supplies discovery signals only; evidence, relevance, atomicity, and duplicate/merge review determine admission. A candidate cannot change the denominator used to score the run that surfaced it.

The recorded human quality pass scanned every critical target across all 10 companies and five queries. It used one company as a calibration example and then downgraded generic registration details, background narrative, absence-of-record claims, weak industry leads, and duplicate formulations to important or relevant. The notes also record parse checks, required-field checks, and duplicate merge-key checks for every company-query group. The versioned review notes do not record the number of individual reviewers, the author of each decision, any independent duplicate annotation, or a formal disagreement-adjudication procedure. We therefore report no inter-annotator agreement and do not claim that the priority labels were independently replicated.

An enrichment stage admitted LLM-assisted candidates. Admission required `action=add`, a positive promotion verdict, one atomic fact, no review or promotion issues, a resolvable evidence path, and no duplicate/merge conflict. Candidates needing evidence or splitting, duplicates, rejected items, unresolved evidence, and non-atomic accepts were excluded. Priority labels use an operational decision rule: *critical* means omission would materially weaken the conclusion for that query, with preference for regulatory filings, announcements, annual reports, prospectuses, official records, strong media, or corroborated sources; *important* supports a key answer judgment; and *relevant* supplies lower-weight context or a lead.

The longer-term maintenance design also aims to bound high-priority-set growth through budgeted selection, demotion, or downgrade rather than unlimited addition. That maintenance mechanism was not used to construct v2: the reported update admitted additions and did not apply a replacement stage. We therefore describe it as future benchmark maintenance, not as part of the reported experimental procedure.

Synthetic worked example. For a credit-and-supplier query, suppose an official regulator notice states that Company X received a material penalty in 2024 for a control failure. The candidate is written as one query-scoped statement—not as the whole notice— with the regulator notice as evidence, detection terms for the event, and a merge key combining event type and date. A second answer-derived paraphrase is merged rather than admitted as another target. The fact is critical only if omitting the penalty would materially alter the credit-risk conclusion; otherwise it is important. An answer sentence alone

cannot supply the evidence record. This example is synthetic and does not disclose any benchmark statement.

The initial review was not designed as a blind annotation experiment. Some admitted candidates were discovered from earlier procedure-index answers, so their candidate provenance was known. This is a property of the update path, not same-run target exposure: evidence and merge gates determine admission, and the source runs retain their earlier frozen denominator.

A.2 Update Review and Evidence Audit

Benchmark updates may use answer-derived discovery signals when an answer reveals a material claim absent from the current snapshot. Admission follows the review gates in Section A.4; a reviewed candidate can enter only a later frozen snapshot.

Evidence references for the admitted subset were checked against the frozen company build manifests and the flattened embedding store. This subset audit supports the availability of those cited source documents; it does not establish complete document-id parity between the two full representations.

A.3 Run Chronology and Main-Row Provenance

The sequential interface ablation and fixed-budget evaluation generated their answers after v2 was fixed, so neither supplied candidates to its own scoring denominator. The main procedure-index row has a different provenance: it is an imported historical answer batch rescored on v2, not a post-freeze rerun. Run identifiers recorded as sources for admitted candidates also occur in that historical batch. Run-identifier overlap establishes provenance overlap, not byte-identical answers, because the benchmark release metadata did not retain answer digests. We therefore treat this as a construction limitation on the main comparison, not as same-run target exposure, and do not present the main row as an independent validation of benchmark construction.

A.4 Review and Apply-Side Validation

Answer text is a discovery signal, not a gold source. A material claim becomes a benchmark fact only after review confirms that it is specific, query-relevant, not a duplicate of any fact with full or partial coverage or a contradiction flag in the same run's coverage audit, and supported by reviewable evidence. The implementation separates three roles: the Answer Fact Recall judge produces the closed-world coverage audit used for scoring; a separate enrichment path extracts answer-side material claims, shapes them into candidate facts, and uses LLM-assisted review and, when available, local evidence supplementation or repair; apply-side validation then blocks candidates whose review output is not safe to carry into the next frozen snapshot.

These checks are construction-quality safeguards, not experimental variables. They include an LLM-assisted candidate review decision, local evidence supplementation or repair for issues with evidence, time, or verifiable elements, resolvable and sufficiently specific evidence routes, fact granularity suitable for answer-level scoring, and duplicate or merge-key checks before the fact enters the next frozen benchmark snapshot. They do not change any current-run recall denominator.

A.5 Unscored-as-Miss Sensitivity

The main embedding and procedure-index arms contain 66 and 126 unscored decisions, respectively. To avoid a missing-at-random assumption, we recompute all four outcomes over the complete frozen target set and assign every unscored decision zero credit. Table 4 shows that all four gaps remain positive and close to the primary estimates. This check changes only the denominator rule; it does not invoke a new judge or repair any decision.

Table 4: Main-comparison sensitivity with unscored decisions counted as misses.

Metric	N	Embedding	Procedure-index	Δ
Critical near-hit	584	0.389	0.510	+0.122
Critical strict	584	0.127	0.250	+0.123
Overall near-hit	3,321	0.265	0.389	+0.124
Overall strict	3,321	0.081	0.187	+0.106

Note: denominators are fixed to all frozen targets in each scope; every unscored decision contributes zero. Embedding has 66 unscored decisions and procedure-index has 126 of 3,321.

A.6 Case De-identification

Case exhibits were assigned Case A–D codenames using a mapping retained outside this release. We removed real company names and aliases, registry and company identifiers, people names, local paths, source filenames, hashes, and identity-bearing directory segments. Mechanism-relevant route categories and approximate magnitudes were retained only where needed to interpret the access pattern. The case descriptions use coarsened regions such as financing, risk, ownership, or cross-module evidence maps rather than raw audit locators.

A.7 Main-Comparison Prompt Contracts

Below are non-executable, instruction-level English translations of the Chinese prompt contracts recorded for the two main arms. They preserve the disclosed instruction order. Bracketed placeholders replace the company and query, local paths, and the proprietary internal traversal sequence; neither prompt exposed the evaluation targets or their target-fact content. The procedure-index arm could consult an existing company-index Answer skill. That skill's internal text is part of the bundled system implementation and is not reproduced in this preprint.

Embedding-arm prompt (translated).

> You are a rapid pre-investment/M&A due-diligence analyst. This run is the embedding-agent MCP baseline. Use only the following embedding-recall tools to obtain evidence:
>
> - `embedding_recall.search`: express the semantic query as a continuous natural-language evidence need, not as keyword filters;
> - `embedding_recall.fetch`: read the bounded evidence text for handles returned by search.
>
> Do not use web search or fetch, local files, repositories, a local index, a private corpus, or the shell. Do not pass a company identifier or filters to search; the policy is cross-company. Complete at least one successful fetch before answering. If evidence is insufficient, say so using the fetched evidence.
>
> **Shared answer task.** Answer the original query as a rapid pre-investment/M&A due diligence. The baselines differ only in available evidence, not in requested analytical coverage. Attach a checkable source name, URL, or permitted evidence handle/path to key claims. Prioritize identity, financial performance, business/products, financing/capital markets, governance, legal/compliance, supply chain/customers, and major risks. Preserve the subject, time or scope, event or metric, and value or conclusion of important facts; do not substitute a different period, entity, or measurement basis. Mark facts not covered by permitted evidence rather than filling them from memory.
>
> **Budget.** At most six searches; default top-$k = 10$, maximum top-$k = 20$; fetch at most 30 documents and 4,000 characters per document.
>
> **Original query.** [ORIGINAL QUERY]

Procedure-index-arm prompt (translated and redacted).

> Use the company-index skill's Answer flow to answer the question. The complete **Shared answer task** block above was repeated verbatim.
>
> **Important constraints.**

- The working directory is an existing read-only index root containing multiple candidate companies: [EXISTING INDEX ROOT]. The target company identifier is not supplied. Locate the subject from the company name or context in the user query.
- You may consult [INTERNAL ANSWER SKILL]. Read, search, and analyze only existing index files. Begin at the located company's existing entry point and follow its navigation to concrete evidence files. [INTERNAL TRAVERSAL SEQUENCE REDACTED.]
- The final answer must contain an "Evidence Sources" section with a Markdown table mapping each key conclusion to concrete, existing evidence files. Do not substitute directories, navigation summaries, wildcards, ellipses, or aggregate counts for those files. When a conclusion depends on multiple records, list the key records separately.
- If the index lacks enough evidence, list the concrete files checked, state the gap, and do not invent paths. Do not build, rebuild, download, collect, or complete the index. Do not create, write, copy, move, or delete files or directories.

User question. [ORIGINAL QUERY]

The statement in the shared block that baselines differed only in available evidence described task semantics: it held requested analytical coverage constant. It did not claim identical full prompt contracts. Both prompts supplied the same original query and excluded benchmark targets; their evidence-access instructions differed, only the procedure-index arm included skill guidance, and the procedure-index prompt imposed no numeric search or fetch budget.

A.8 Corpus and System Configuration

The procedure-index and flattened embedding representations were derived from the same underlying local source environment. The shared frozen snapshot contains 29,688 materialized Chinese-language evidence documents, distributed across the 10 companies as summarized in Table 5. Document counts and length statistics are computed from the flattened vector-store document file and validated against its build manifest. This release does not include a complete endpoint-level record of attempted, failed, or rejected upstream collection operations; the table therefore characterizes the realized snapshot rather than a collection-success rate.

Table 5: Per-company statistics of the shared frozen snapshot, computed from the flattened vector-store document file.

Item	Value
Documents	29,688
Documents/company (min / median / max)	598 / 2,722 / 5,966
Characters/document (median / p90 / p95)	556 / 2,717 / 5,442
Documents over 4,000 chars	1,908 (6.4%)
Documents over 8,000-char input cap	1,079

The main-comparison system configuration—shared answer-model name and harness, separate run generations, arm-specific representation, tools, prompts, and budget, unpinned endpoint-default decoding—is specified in the main text; the recognizer is a separate closed-world coverage model.

A.9 Legacy Public-Web Reference Rows

Table 6 reports the legacy web-only and hybrid rows on the same frozen benchmark. They are historical imported baselines whose discovery surface includes the public web, so they serve as reference context rather than paired arms: they were not rerun under the main protocol and are excluded from all planned contrasts. Their strong critical near-hit recall indicates what public-web discovery can reach; the controlled question of this paper is what local evidence organization contributes over the same local corpus.

Table 6: Legacy public-web reference rows on the same frozen benchmark. Historical imported baselines with a different discovery surface; not paired reruns and excluded from all planned contrasts.

Metric	Hybrid-Web	Web-Only
Runs	50	50
Critical NH	0.664 [0.614, 0.726]	0.692 [0.632, 0.761]
Critical strict	0.288 [0.243, 0.345]	0.317 [0.265, 0.384]
Overall NH	0.405 [0.361, 0.450]	0.405 [0.367, 0.447]
Overall strict	0.139 [0.118, 0.160]	0.138 [0.122, 0.155]

Note: NH = near-hit recall; strict = fully expressed fact recall; recall denominators include scored facts only; CI = company-cluster bootstrap percentile interval over 10 companies.

A.10 Recognizer Reliability

To bound serving-side scoring noise, we first re-scored every main-comparison answer with the identical DeepSeek recognizer configuration (same prompt version and configuration hash, temperature 0). Agreement was high and symmetric across arms, and all four planned main contrasts remained at $p \leq 0.004$. We then repeated the same closed-world scoring pass with an independent GLM-5.2 judge. Table 7 reports the cross-judge agreement: both arms contain all 50 answer blocks, all 3,321 fact decisions join by fact key, and no answer is skipped. The GLM-5.2 pass is more conservative on strict recall, but the difference is not arm-differential and all planned contrasts remain at $p \leq 0.002$.

Table 7: Recognizer cross-judge agreement on the main-comparison answers (DeepSeek primary judge vs. GLM-5.2 secondary judge).

Arm	Status	Coverage (κ)	Strict (κ)	Near-hit (κ)
Embedding	98.0%	90.1% (0.78)	96.9% (0.75)	91.2% (0.78)
Procedure-index	96.2%	89.1% (0.81)	95.2% (0.83)	92.1% (0.84)

Note: per-fact decision agreement between the two scoring passes over 3,321 joined decisions per arm. Between the passes no pooled recall estimate moves by more than 0.057, and the largest planned-contrast p-value is 0.002.

These checks bound recognizer sensitivity; they do not establish agreement with human judgment. The closed-world contract provides an additional mechanical check on positive decisions: every covered or partially covered fact must quote a verbatim span of the answer, and spans are validated against the answer text. The reported validation stops at the LLM cross-check; its result is a closed-world fact-coverage measurement, not a general human-preference judgment.

A.11 Sequential Interface Ablation Estimates

Table 8 reports point estimates and intervals for all four cumulative profiles. The main paper uses Figure 2 for the ordered pattern and Table 3 for the adjacent H/N/S package contrasts.

A.12 Budget Sensitivity and Stopping Protocols

The original fixed-budget paired evaluation runs every case through the same two-stage protocol: an *evidence-collection* stage followed by an *answer-synthesis* stage. Evidence collection uses each arm's usual evidence-access tools at a shared $0.40 budget threshold.

The $0.40 threshold was selected through pilot calibration to create a binding evidence-access regime in both arms. The calibration criterion was operational: evidence collection should encounter model-API spending pressure in both interfaces, requiring each agent to prioritize what to inspect. In the published runs, 48 of 50 embedding runs and all 50 procedure-index runs ended evidence collection with `budget_exhausted`; the other two embedding runs ended `completed`. Both are protocol-legal graceful endings.

Table 8: Sequential interface ablation estimates. Each cell reports pooled recall with a 95% company-cluster bootstrap CI.

Arm	Critical	Overall
Near-hit recall		
Flat (000)	0.382 [0.307, 0.467]	0.290 [0.239, 0.342]
+Hierarchical evidence (100)	0.418 [0.369, 0.483]	0.336 [0.284, 0.385]
+Navigation documents (110)	0.551 [0.488, 0.629]	0.399 [0.362, 0.438]
+Skill guidance (111)	0.558 [0.491, 0.637]	0.409 [0.360, 0.460]
Strict recall		
Flat (000)	0.125 [0.090, 0.171]	0.100 [0.077, 0.126]
+Hierarchical evidence (100)	0.180 [0.136, 0.232]	0.136 [0.107, 0.169]
+Navigation documents (110)	0.217 [0.185, 0.247]	0.148 [0.131, 0.164]
+Skill guidance (111)	0.240 [0.206, 0.280]	0.161 [0.138, 0.182]

Note: all arms contain 50 runs; recall denominators include scored facts only.

Table 9: Fixed-budget boundary check: the overall near-hit difference remains positive and suggestive at this operating point; the other three planned contrasts are not significant. Cells report pooled fact recall with 95% company-cluster bootstrap CIs.

Metric	Embedding	Procedure-index	Difference	p
Critical near-hit	0.373 [0.329, 0.411]	0.423 [0.356, 0.497]	+0.050	0.293
Critical strict	0.145 [0.116, 0.170]	0.141 [0.104, 0.177]	-0.004	0.906
Overall near-hit	0.254 [0.227, 0.284]	0.297 [0.259, 0.338]	+0.044	0.059
Overall strict	0.084 [0.073, 0.098]	0.097 [0.078, 0.118]	+0.012	0.324

Note: Both arms contain 50 runs; difference is Procedure-index minus Embedding; recall denominators include scored facts only; planned contrasts use exact two-sided company-cluster permutation p-values; p-values are nominal and unadjusted; bold marks nominal $p < 0.05$; protocol eligibility and scored counts are 50/50 for both arms.

This calibration defines one experimental operating point; it does not establish \$0.40 as a deployment-optimal budget or as an optimum on a cost–recall curve. The threshold applies only to evidence-collection model-API spending during search and navigation. It is enforced at model-call boundaries, so the final atomic call can take realized spending above the nominal \$0.40 threshold. It therefore does not guarantee identical realized spending, retrieval-action counts, evidence volume, or total system cost.

Answer synthesis resumes the evidence-collection conversation history as a fork and synthesizes the answer with an identical prompt and all tools disabled; answer-synthesis spending is outside the ceiling, and no notes files or workspace artifacts carry over. Evidence collection remains arm-specific in representation, tools, prompts, action counts, evidence volume, and observed history. Differences in action count are therefore part of the complete-interface comparison rather than a matched experimental quantity.

Eligibility requires a legal evidence-collection stop, a non-empty answer-synthesis response, and zero answer-synthesis tool calls; all 50 cases per arm pass.

Natural-stop sweep. The natural-stop extension retains the same two-stage protocol while varying the nominal evidence-collection cap. The procedure-index and embedding curves share 10 caps: \$0.40, \$0.50, \$0.60, \$0.70, \$0.80, \$0.90, \$1.00, \$1.30, \$1.60, and \$2.00. This was an adaptive rather than preregistered fixed grid: a completed coarse grid at \$0.40, \$0.70, \$1.00, \$1.30, \$1.60, and \$2.00 was enriched with \$0.50, \$0.60, \$0.80, and \$0.90 after curve review. The original \$0.40 paired check above comes from a separate run generation and retains its own planned tests; its estimates are not substituted with the

Table 10: Embedded forced-continuation sweep. Recall is pooled over scored facts; spending and calls are means over 50 cases. The $1.20 point is forced-only and is not plotted in Figure 3.

Cap ($)	Stage A spend ($)	Search	Fetch	Strict	Near-hit
0.40	0.440	9.20	2.90	0.082	0.295
0.50	0.550	10.36	3.80	0.088	0.305
0.60	0.645	11.62	4.64	0.094	0.306
0.70	0.749	12.68	4.94	0.096	0.340
0.80	0.850	13.94	5.20	0.096	0.356
0.90	0.941	14.90	5.54	0.105	0.346
1.00	1.048	15.58	6.28	0.097	0.349
1.20	1.254	16.90	6.88	0.110	0.366
1.30	1.350	18.88	7.88	0.101	0.360
1.60	1.654	20.78	8.76	0.107	0.358
2.00	2.046	23.52	10.00	0.110	0.390

sweep's $0.40 point.

Under natural stopping, an ordinary legal end turn may terminate evidence collection before the nominal cap. The embedding sweep removes fixed search-call and fetch-document limits, but the agent may still decide to stop. Stopping-status counts across the grid are reported in Section 5.3. The x-axis therefore represents a nominal cap, not an experimentally assigned amount of realized acquisition effort.

Forced-continuation follow-up. The `forced-budget-evidence-collection-resume-answer-v1` protocol applies only to the embedding arm. Evidence collection may not produce the answer: after each ordinary end turn it must resume the same session, add at least one newly fetched canonical evidence item, and continue until the cumulative budget terminal. Answer synthesis then forks the same session once with all tools disabled. The procedure-index arm was not run under this protocol; its natural-stop curve is repeated in Figure 3 only as a descriptive reference.

The formal forced sweep contains 11 nominal caps: $0.40, $0.50, $0.60, $0.70, $0.80, $0.90, $1.00, $1.20, $1.30, $1.60, and $2.00. Each point contains all 50 cases, one generation per case, and zero unscored facts. The $1.20 point is excluded from the three-series figure because neither natural-stop curve has a matching point; Table 10 reports the complete forced grid. The endpoint contrasts reported in the paper were selected before the formal forced sweep was run.

All sweep point estimates pool scored facts over the 50 cases. Intervals use a 95% company-cluster bootstrap with 10 company blocks. They preserve within-company dependence among query types but do not measure generation variance. Nominal cap, realized model spending, search and fetch actions, and the quantity or novelty of acquired evidence remain distinct quantities.