

SCIENTIFIC WHITE PAPER

External Knowledge for Scientific Agents

Testing one-shot external knowledge compilation—
and what the experiments changed

PUBLISHED

September 2026

ORGANIZATION

Beakr Inc.

WEB

thebeakr.com

AT A GLANCE

Compilation can improve reuse, cost, and performance.

The experiments do not establish that one task-agnostic representation is sufficient for unknown future tasks—so the architecture stays hybrid.

RESEARCH QUESTION

Can organizational information be compiled once into a general-purpose external representation—analogous to a foundation model's internal knowledge, but mutable, source-linked, and shared?

FOCUSED BENCHMARKS • BEAKR + KB VS BEAKR + RAW SOURCES, PAIRED ON THE SAME QUESTIONS

ENTITY RESOLUTION

-77%

Query cost

Accuracy unchanged. 18 questions over 90 files.

EVIDENCE SYNTHESIS

+23 pts

Correctness

Cost unchanged within uncertainty. 48 questions, 31 FDA drug labels.

TEMPORAL CHANGE

+7 pts

Strict accuracy

Query cost 52% lower. 30 questions, three corpus states.

CONTRADICTIONS

+33 pts

Recovery

Query cost 71% lower. 12 questions, 12 dossier files.

REALISTIC & EXTERNAL SETTINGS

AMRBENCH

95.2%

Answer correctness (59/62)

Highest observed correctness and lowest mean latency (60.8 s) across retained historical runs; descriptive.

HARVEY LAB

+4.0 pts

Criteria score, same Codex agent

67.7% → 71.7% on Harvey's 250-task firm-knowledge benchmark with the Beakr KB added.

PBH COMPOUNDING PILOT

+31.4 pts

Four-task synthesis mean

41.5 → 72.9 after reviewed corrections were captured; answering cost 62% lower. Exploratory.

Values are rounded for clarity; caveats for each result appear in the sections that follow. Several of the focused benchmarks' paired 95% bootstrap intervals are wide, so those experiments should be read as mechanistic proofs of concept rather than a general performance ranking.

01 —

Why external knowledge matters now

CONTENTS

01	Why external knowledge matters now	3
02	The Beakr knowledge system	3
03	An evaluation suite for stateful knowledge	4
04	Focused curated benchmarks: tests of the KB design assumptions	5
05	External system comparison	5
06	From controlled assumptions to realistic environments	6
07	From static knowledge to compounding knowledge	8
08	Conclusion and future directions	9
	References	9

Scientific organizations increasingly use foundation models to execute well-defined tasks. Organizational knowledge, however, has different requirements: it is private, continuously changing, distributed across systems, frequently contradictory, and subject to provenance and access controls. Encoding this state only in model weights couples knowledge updates to training, obscures the origin of individual claims, and makes the knowledge difficult to transfer between models. Supplying every source document directly at inference time preserves detail but repeatedly incurs the cost of locating and interpreting the same information; long context windows also do not guarantee reliable use of every relevant fact^{1,2}.

Foundation models already compress broad corpora into internal, parametric knowledge that is reusable across tasks. For organizational knowledge, we considered an external analogue: a persistent representation that remains independent of any single model, can be updated and governed, and preserves provenance. Recent external-memory systems report improvements on specific downstream and long-term-memory benchmarks, demonstrating end-to-end utility. Such results, however, do not isolate whether a compiled representation itself remains sufficient when downstream tasks are unknown. Our initial hypothesis was that one task-agnostic compilation could support broad future use. In the focused studies, we held the agent harness and questions fixed while varying access to the knowledge base (KB) versus the raw sources; broader experiments then tested end-to-end behavior in more realistic settings. The results show that compilation can improve reuse, cost, and performance, but do not establish that one task-agnostic representation is sufficient for unknown future tasks.

02 —

The Beakr knowledge system

Beakr separates compilation, persistent state, and use. Its knowledge base (KB) is implemented as a versioned, source-linked wiki whose pages represent project entities, relationships, chronology, claims, and provenance. Compiler extracts entities, relationships, chronology, and source-linked claims from documents, messages, scientific data, and databases, then structures and synthesizes proposed KB changes. Review validates changes before commit: the first creates T0 and later commits create T1, T2, and subsequent versions. A user, whether a person or an agent, queries the current state through Ask. Capture branches from that interaction, recording evidence-backed corrections and insights before returning them to Compile. A project profile guides Compile and provides context to Ask.

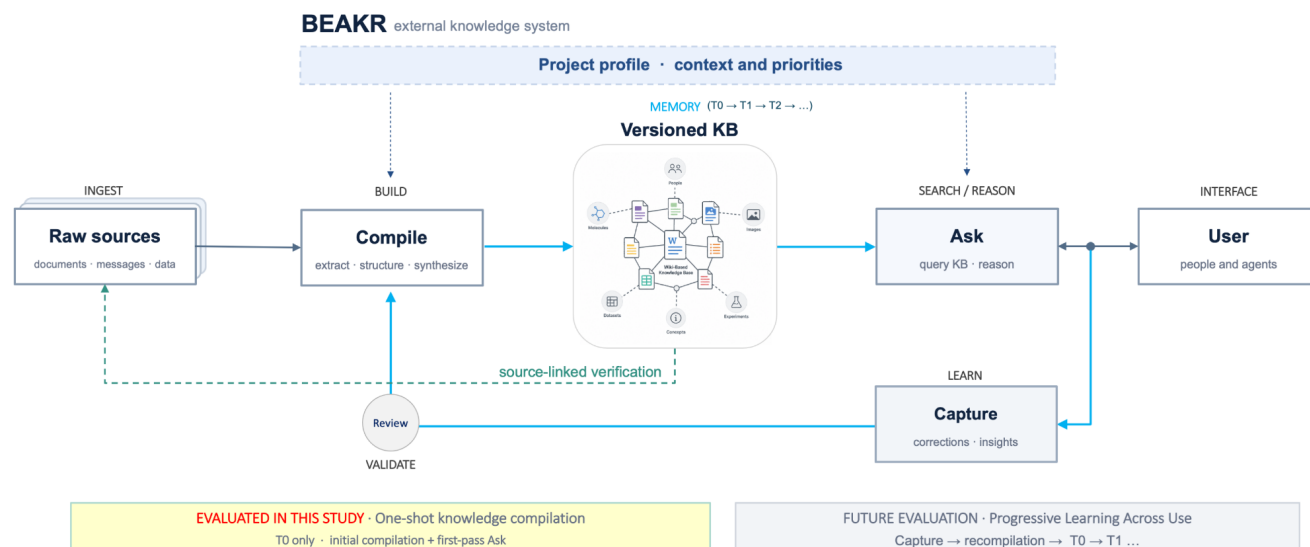


FIGURE 01 The Beakr external knowledge system. The primary path creates and queries a Versioned KB. Capture branches from the interaction with Ask and returns evidence-backed updates to Compile. The dashed path denotes source verification. Most controlled results in this report evaluate T0; the PBH compounding pilot separately tests reuse after reviewed corrections are captured into persistent project knowledge.

03 —

An evaluation suite for stateful knowledge

TWO MODES

Frozen T0 ablations

Raw sources · KB-only ·
KB + Raw sources

Progressive trajectories

Updates from earlier use
shape later state

Beakr's Eval Studio evaluates the knowledge system, not only its final answers. It can freeze the KB at a T0 snapshot for matched Raw sources, KB-only, and KB-plus-Raw sources ablations, or run progressive trajectories in which accepted updates arising from earlier use create the state presented to later questions. Lightweight adapters keep questions and target-agent settings constant while measuring answer correctness, evidence precision and recall, source recovery, and faithfulness alongside latency, tool use, tokens, and cost. Results are interpreted within task, and repeated T0 builds and Ask runs separate representation effects from execution variance.

03.1

Evaluation strategy

The evaluation separates two questions. The first section, *Focused curated benchmarks: tests of the KB design assumptions*, isolates one KB design assumption at a time while holding the questions and Beakr harness fixed. Within each benchmark, Beakr + KB is compared with the same system given direct access to raw sources, so the paired difference estimates the contribution of the compiled KB. The second section, *From controlled assumptions to realistic environments*, tests whether those conclusions hold when corpora are heterogeneous, future questions are unknown, or the benchmark originates outside Beakr.

Focused curated benchmarks: tests of the KB design assumptions

TABLE 01 Four KB design assumptions, each isolated in its own focused benchmark.

KB ASSUMPTION	KNOWLEDGE VARIABLE	BEAKR: KB VS RAW SOURCES	INTERPRETATION
A1 Stable identities can be resolved once and reused.	Entity resolution. 18 questions over 90 files describing 18 evolving projects with aliases and name changes.	Accuracy unchanged; query cost 77% lower.	Resolve each identity once, then reuse it.
A2 A compiled KB can support evidence synthesis.	Evidence synthesis. 48 questions across 31 FDA drug-label files.	Correctness 23 points higher; cost unchanged within uncertainty.	Pre-organized evidence can improve multi-document answers.
A3 A compiled KB can retain changing facts.	Temporal change. 30 questions across three corpus states containing 100, 108, and 116 clinical-trial files.	Strict accuracy 7 points higher; query cost 52% lower.	The KB can help carry scoped updates forward.
A4 Explicit conflict structure can expose contradictions.	Contradictions. 12 questions across 12 scientific dossier files containing mutually inconsistent constraints.	Recovery 33 points higher; query cost 71% lower.	Structured conflicts are easier for agents to recognize.

Paired mean differences compare Beakr + KB with Beakr + Raw sources on the same questions. Values are rounded for clarity. Paired 95% bootstrap intervals are omitted here for legibility; several are wide, so these experiments should be interpreted as mechanistic proofs of concept rather than a general performance ranking.

External system comparison

To test whether the focused findings extended beyond the controlled Beakr + Raw sources baseline, we compared Beakr + KB with two external KB systems (OpenWiki and Chroma) and two frontier raw-source baselines (Claude and Codex) on the same corpora and questions. Because the systems differ in answering model, retrieval interface, and cost accounting, these are paired end-to-end comparisons—not an isolated test of KB representation or a universal ranking.

Results were task-dependent. Entity resolution was saturated, with Beakr + KB generally reducing query cost. Beakr improved evidence synthesis relative to Chroma and contradiction recovery relative to Chroma and Claude, while Chroma achieved higher temporal-change accuracy at substantially greater cost. Several paired intervals crossed zero, and OpenWiki’s temporal result was affected by a citation-format mismatch. Overall, compiled external state can improve efficiency

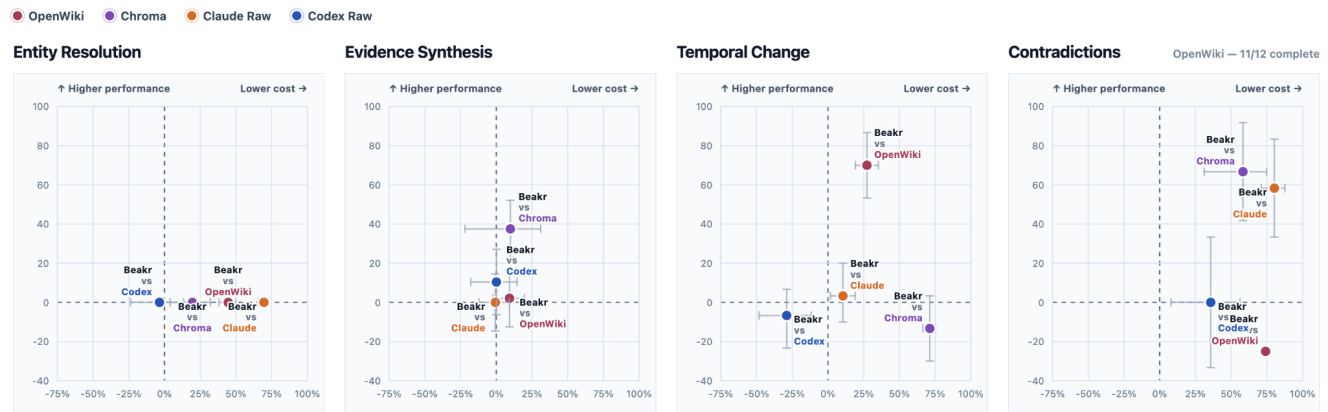


FIGURE 02 Paired performance and query-cost comparisons of Beakr + KB against external KB systems and frontier models with raw-source access. Each point compares Beakr + KB with one external KB system or frontier raw-source baseline on the same benchmark questions. Vertical position is Beakr + KB performance minus comparator performance (percentage points); horizontal position is query-cost savings relative to the comparator. Up and right favor Beakr + KB; whiskers show paired 95% bootstrap confidence intervals. Query cost includes target-agent API-equivalent inference only; one-time KB construction is excluded, and Chroma cost is token-estimated.

and task-specific accuracy, but its value depends on preserving the task’s unit of truth.

06 —

From controlled assumptions to realistic environments

The focused benchmarks above isolate situations in which the relevant unit of knowledge is known in advance. We next tested whether a compiled representation remains useful in more realistic settings, where corpora are heterogeneous and future questions are not known during compilation.

06.1

59/62

ANSWER CORRECTNESS

95.2% on the shared questions, with the lowest mean latency (60.8 s). Descriptive: historical runs.

AMRBench: one-shot compilation under realistic noise

AMRBench provides an end-to-end stress test of the stronger one-shot hypothesis: can a task-agnostic compiler create a reusable external representation before future questions are known? Its deliberately authored antimicrobial-resistance lab world contains 86 files spanning Slack-style threads, tabular records, prose reports, and six frozen public sources from EUCAST, NCBI, and Europe PMC. Twenty-nine decoy files force provenance, contradiction, and temporal questions to distinguish the correct record from plausible alternatives.

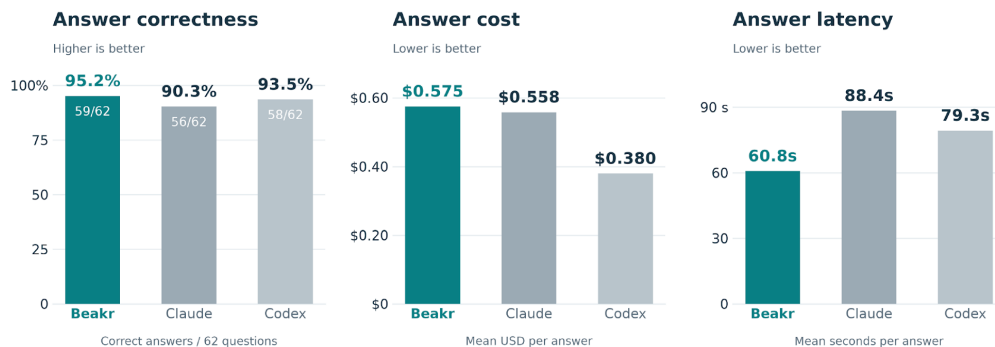
On the 62 questions shared across retained historical runs, Beakr achieved the highest observed answer correctness (59/62; 95.2%) and the lowest mean latency (60.8 s), compared with Claude (56/62; 90.3%; 88.4 s) and Codex (58/62; 93.5%; 79.3 s);

FIGURE 03 Observed AMRBench performance across retained historical runs. On 62 shared questions, Beakr achieved the highest observed answer correctness and lowest mean latency; Codex had the lowest mean query cost. Runs were historical rather than contemporaneous matched reruns, and costs exclude KB construction and infrastructure; the comparison is descriptive rather than causal.

KLB / OBSERVED QUALITY, COST AND SPEED

Beakr: highest correctness and lowest latency

Common 62-question comparison | One historical run per agent



Identical question prompts and reference answers for the 62 shared cases; the Beakr 74-case run is restricted to that overlap.

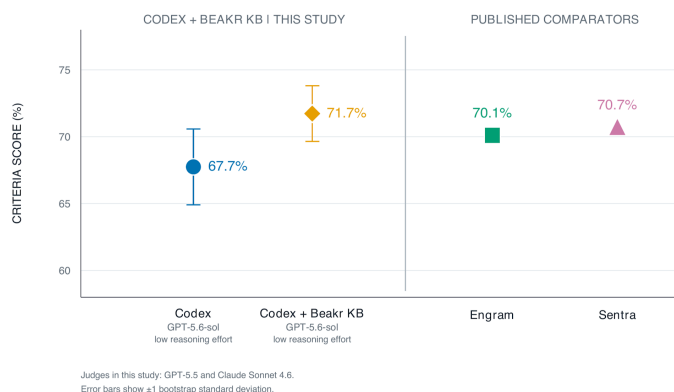
Historical runs differ in configuration and generation. One answer per question; differences are descriptive.

Answering costs exclude KB construction and infrastructure. Answer correctness is separate from strict all-gate success.

FIGURE 04 Beakr improves the same Codex agent on Harvey LAB. Across Harvey's 250-task firm-knowledge benchmark, adding the Beakr KB increased mean criteria score from 67.7% to 71.7% (+4.0 points). Engram and Sentra are published results from separate studies and are included as contextual, not paired, comparisons.

Harvey LAB firm-knowledge criteria score

Reported share of rubric criteria passed across the 250-task benchmark



Codex had the lowest mean query cost. Because these were not contemporaneous matched reruns, Figure 3 is descriptive: it shows that Beakr reached a favorable observed correctness–latency frontier, not that the KB effect has been causally isolated.

06.2

External benchmark: generalization beyond Beakr-authored corpora

We then asked whether the same conclusions transferred beyond Beakr-authored corpora. Harvey LAB provides an external performance test.

+4.0 pts

SAME CODEx AGENT

Criteria score 67.7% → 71.7% with the Beakr KB added.

HARVEY LAB

On Harvey's 250-task firm-knowledge benchmark³, adding the Beakr KB to the same Codex model increased the criteria score from 67.7% to 71.7% (+4.0 points). The Beakr-assisted result also exceeded the published Engram (70.1%) and Sen-tra (70.7%) scores. Because those published values come from separate studies, the strongest causal evidence is the paired Codex ablation: Beakr improved the same agent on the same benchmark while reaching the range of leading published systems.

07

+31.4 pts

FOUR-TASK MEAN

41.5 → 72.9 under a paired frozen-contract regrade; answering cost \$7.51 → \$2.84.

From static knowledge to compounding knowledge

The PBH (primordial black hole) compounding pilot is a probe on whether reviewed scientific corrections can become reusable knowledge across episodes. Its 42 synthetic research artifacts are grounded in three astrophysical workflows—particle-transport outputs, accretion-response archives, and cosmological-inference tables. The ten initial cases require distinguishing physical effects from simulation or export errors, checking numerical and provenance assumptions, and deciding whether a scientific result should be released, corrected, or withheld. Beakr answered those cases in one shared project, received domain-expert suggestions, and captured the reviewed findings into the KB.

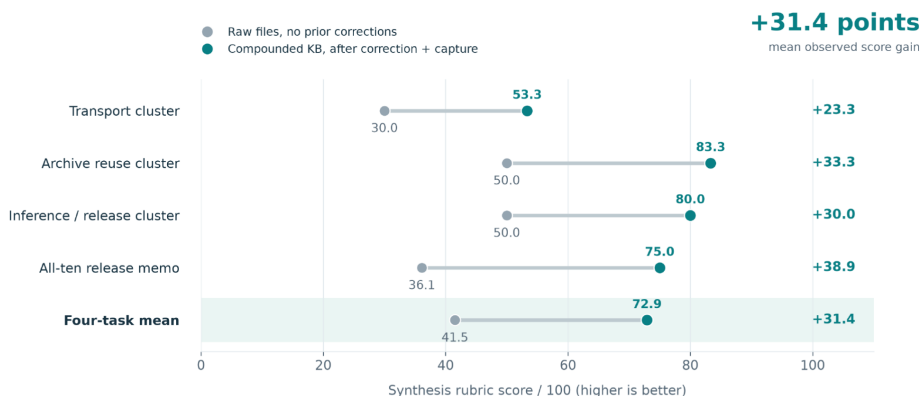
We then tested increasingly broad synthesis in fresh conversational threads which precluded access to the corrected historical threads: three topic-cluster

FIGURE 05 PBH compounding pilot. The same four synthesis prompts were answered in fresh threads using either all 42 raw artifacts with no prior corrections or a compounded KB containing reviewed corrections with no raw-source access. Scores use a paired frozen-contract regrade; because correction history and access mode both change, the comparison is exploratory rather than a causal KB ablation.

PBH COMPOUNDING PILOT / BEAKR ONLY

Retained knowledge supported stronger synthesis

The same four dendrogram prompts, answered in fresh threads under two knowledge conditions



One answer per condition per task; paired frozen-contract regrade (2-3 judgment passes per answer).

Exploratory contrast: correction history and file/KB access both change. Higher rubric scores do not establish safe release decisions.

questions combined findings from related cases, and an all-ten release memo required a portfolio-wide decision and repair plan. The compounded-KB condition used the accumulated project knowledge without raw-source access; the comparison condition received all 42 raw artifacts but no correction history or KB. Under a paired frozen-contract regrade, the four-task mean increased from 41.5 to 72.9 (+31.4 points); all four tasks improved, and the all-ten memo gained 38.9 points. Answering cost fell from \$7.51 to \$2.84 (62% less), excluding prior learning and KB preparation. Because correction history and access mode changed together, this is an end-to-end test of Beakr’s compounding mechanism, not a pure KB ablation.

08 —

Conclusion and future directions

The experiments support external knowledge as a reusable scientific substrate, but they do not establish the stronger claim that a semantically sufficient organizational memory can be compiled once for unknown future tasks. Initial compilation should establish a general structural backbone—entities, relationships, chronology, and provenance—while richer synthesis develops as actual questions reveal what matters.

The project profile provides a task-relevant prior: goals, scientific context, focus areas, and key entities guide Compile toward information likely to matter and help Ask interpret questions and frame answers, without requiring future questions to be specified in advance. The production architecture should therefore remain hybrid: the KB provides evolving shared state and navigation, while raw sources remain authoritative for verification, exceptions, and novel questions.

Future work should train compiler and memory models across diverse corpora, profiles, and tasks to learn what information must survive compilation; develop routing policies that select the KB, raw sources, or both according to the query’s required unit of truth and the KB’s evidentiary coverage; and test whether the PBH pilot’s compounding gains persist across longer task sequences without eroding evidence fidelity.

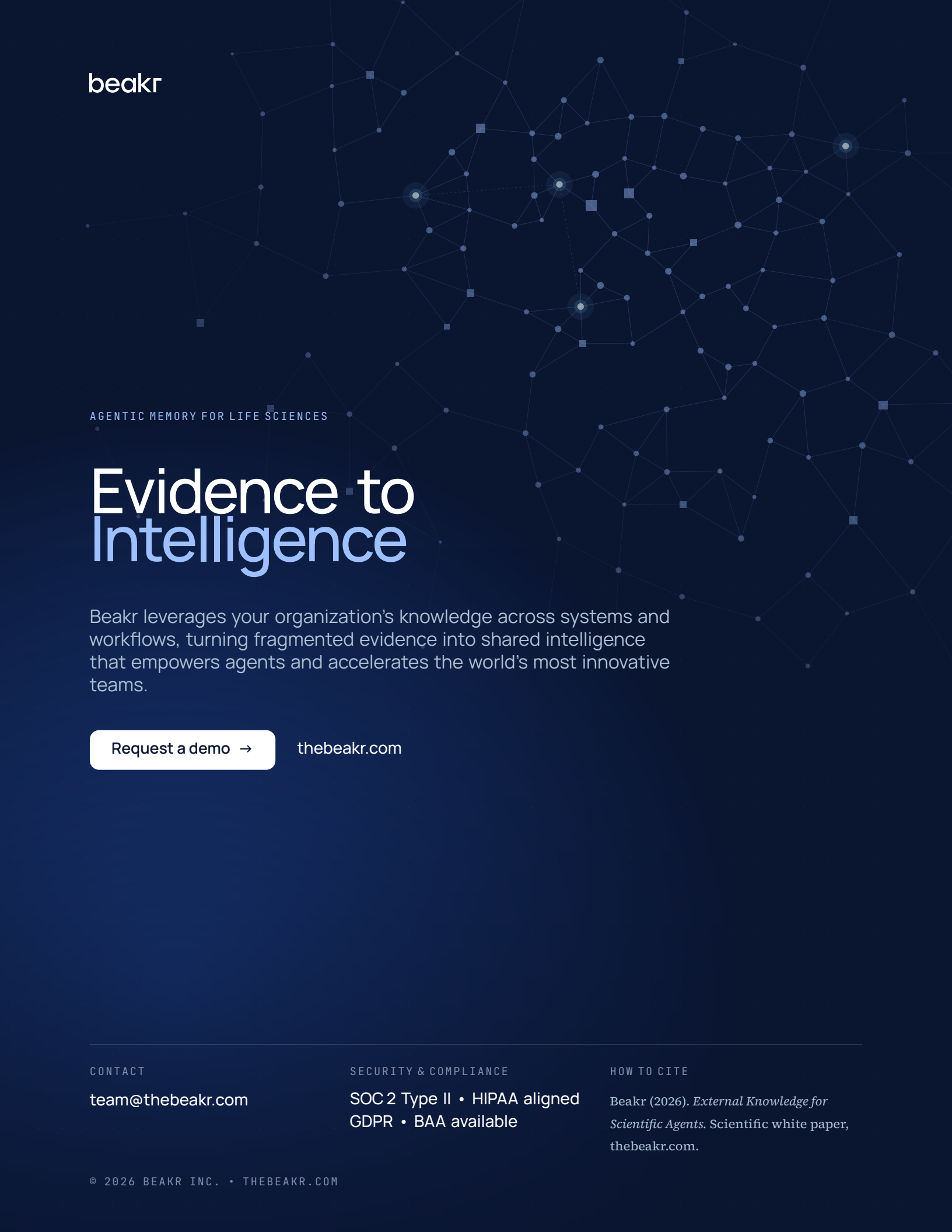
Thus the architecture should be semantic-first, not semantic-only: preserve loss-less records where precision or completeness matters, and route to raw evidence whenever the KB cannot support the required unit of truth.

“

Semantic-first,
not semantic-only.

References

- 1 Lewis, P. et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv:2005.11401*.
- 2 Liu, N. F. et al. (2024). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12, 157–173.
- 3 Pereyra, J. et al. (2026). LAB: Law Firm Knowledge. Harvey. <https://www.harvey.ai/blog/legal-agent-bench-law-firm-knowledge>.



beakr

AGENTIC MEMORY FOR LIFE SCIENCES

Evidence to Intelligence

Beakr leverages your organization's knowledge across systems and workflows, turning fragmented evidence into shared intelligence that empowers agents and accelerates the world's most innovative teams.

Request a demo →

thebeakr.com

CONTACT

team@thebeakr.com

SECURITY & COMPLIANCE

SOC 2 Type II • HIPAA aligned
GDPR • BAA available

HOW TO CITE

Beakr (2026). *External Knowledge for Scientific Agents*. Scientific white paper, thebeakr.com.