Docket

DOCKET / RESEARCH & PRACTICAL GUIDANCE

# AI Research Agent Benchmark for Acquisition Target Research

> Build a more reviewable acquisition pipeline with Docket. Our deal-origination software and managed research help private equity teams screen companies against a mandate and understand the evidence behind each finding.

Published by Docket · 2026-09-19 · ai research agent benchmark · ai agent evaluation framework · acquisition target research · private equity ai · deal sourcing tools · evidence quality · human adjudication · agent benchmarks · research operations

Original article: https://docket.capital/articles/benchmark-ai-agents-acquisition-research



### In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes: From Demo to Decision
4. Build the Benchmark Pack
5. Implementation Considerations and Process Changes
6. Scoring and Human Adjudication
7. Data Analysis and Evidence
8. Procurement Decision and Release Process
9. Implications and Future Directions
10. Frequently Asked Questions (FAQs)
11. Conclusion

# Executive Summary

An acquisition-research benchmark should decide whether a system is safe and useful for a specified screening task, not whether it writes an impressive report. The evaluation unit is **company x research question x collection date**. Each case should have a frozen mandate, frozen source corpus, allowed tools, exact system configuration, gold status, acceptable variants, and a risk tier. This design reflects the fact that web resources change over time (www.w3.org) and that official registers themselves can be incomplete or lightly checked (www.gov.uk). A persuasive narrative is therefore not the scoreable object.

The benchmark pack in this report uses **stratified cases**, **two human reviewers**, an explicit allowed-answer set, a hidden final test split, and at least **three repeated runs** per system as a practical starting point. Repetition matters because model outputs are generally sampled nondeterministically (nvlpubs.nist.gov), while a current Microsoft evaluation guide independently recommends at least three runs before conclusions are drawn (learn.microsoft.com). Three runs do not prove stability, but they reveal obvious variance. Preserve every answer, citation, retrieved page, tool call, prompt hash, token cost, elapsed time, and review minute.

Release decisions should be field-specific. **Identity** can pass only with extremely high precision because a wrong entity contaminates every downstream answer. **Ownership** and disqualifying mandate facts require fresh human review. **Revenue estimates** should remain estimates with disclosed methods and must never be promoted to verified facts. Score answer correctness, citation faithfulness, completeness, sufficiency, source authority, temporal validity, identity resolution, abstention quality, and prohibited-inference rate separately. NIST's current project demonstrates faithfulness, completeness, and sufficiency probes (www.nist.gov), but its repository says more validation of probes, rubrics, and model-judge performance is needed (github.com). Models may triage disagreements, not settle high-risk facts.

The economic denominator is **accepted answers**, not generated words or records attempted. Report total run plus reviewer cost divided by accepted answers, alongside p50 and p90 elapsed time and review minutes. A system passes only when every non-negotiable gate clears on the hidden set and no failure cluster is

concealed by an overall average. NIST recommends paired case-level comparisons and reporting cost with performance (nvlpubs.nist.gov). A small pilot can support a limited deployment decision, but it cannot establish universal system superiority or deal outcomes.

## Introduction and Background

Private-equity origination teams increasingly compare three operating models: a general **AI research agent**, a purpose-built vendor platform, and a human-supported research workflow. The procurement question is narrow: which system can answer this mandate's ownership, identity, fit, size, and contact questions with evidence at an acceptable error rate? Generic AI agent evaluation frameworks help, but acquisition-target research adds entity ambiguity, source conflict, time-sensitive ownership, absence-based answers, and uneven private-company footprints.

The benchmark must therefore begin with a decision and explicit release gates. NIST says objectives should specify the construct measured and the intended uses of results (nvlpubs.nist.gov). It also cautions that automated benchmarks cannot meet every evaluation objective (nvlpubs.nist.gov). As of September 2026, **NIST AI 800-2 remains an Initial Public Draft**, not a final standard (www.nist.gov). This report uses its practices as authoritative draft guidance, not certification criteria.

Docket is a direct provider in this category. Its public description says Triage checks identity and fit, Scout collects answers, and Audit reviews evidence. Its existing conceptual guide already recommends a held-out, human-scored sample and per-question measures. The present report does something different: it specifies an executable bake-off pack. Docket appears in the contender table as required for a direct provider, but no score is assigned because no controlled evaluation has been run here.

## Key Changes: From Demo to Decision

### Change 1: score atomic cases, not narratives

The minimum scoreable record is:

- **Company key:** a stable internal identifier plus observed names and domains.
- **Research question:** one mandate field with an answer schema.
- **Collection date:** the date on which the evidence was available.
- **Gold status:** supported value, justified not-found, conflicting, not applicable, or unresolved.
- **Allowed variants:** normalized names, units, date tolerances, and acceptable categorical equivalents.
- **Evidence:** source URL, retained excerpt, source date, access date, and authority class.

This unit prevents a fluent report from hiding a wrong field. It also respects source temporality. W3C provenance guidance models changing states as multiple entities with their own identifiers (www.w3.org), while the Memento protocol states that an original resource's state may change over time (www.rfc-editor.org). A correct ownership answer collected last year can be wrong today without either observation being poorly researched.

## Change 2: separate correctness, support, and coverage

A single accuracy percentage is insufficient. A benchmark should measure at least these dimensions independently:

- **Answer correctness:** does the normalized answer match an allowed gold value?
- **Citation faithfulness:** does the cited span support the claim? NIST defines the probe in exactly this claim-source relationship (www.nist.gov).
- **Completeness:** does the answer preserve relevant qualifications? (www.nist.gov).
- **Sufficiency:** does the source carry the evidentiary burden? (www.nist.gov).
- **Coverage:** how many required questions received acceptable answers?
- **Abstention quality:** did the system decline when evidence was absent or conflicting?
- **Identity resolution:** did all evidence refer to the intended company?
- **Temporal validity:** was the answer current at the collection cutoff?
- **Prohibited inference rate:** how often did the system turn a clue into an unsupported fact?

This separation makes tradeoffs visible. A conservative system may have high claim precision but low coverage. A broad system may answer more questions while attaching weak sources. Neither profile is adequately represented by one composite score.

## Change 3: compare controlled workflows

*Table 1* defines the contender classes and the parity controls needed before scores can be compared.

| Contender | Operating model | What must be held constant | Decision interpretation |
|---|---|---|---|
| **Standalone AI research agent** | Model plus prompt, tools, and retrieval workflow | Mandate, corpus or web cutoff, paid-source access, time limit, output schema, and review policy | Tests the configured system, not the base model alone |
| **Vendor platform** | Packaged research product with proprietary orchestration or data | Same required questions, dates, geography, evidence standard, and allowable source classes | Feature breadth does not compensate for a failed evidence gate |
| **Docket** | Direct provider using Triage, Scout, and Audit stages; self-serve or managed delivery | Same hidden cases and independent adjudicators; disclose any retained-page or workflow advantage | Treat as one candidate workflow; this report publishes no score |
| **Human-supported workflow** | Analyst research, possibly with AI assistance | Same mandate, source access, time budget, and accepted-answer definition | Provides an operational comparator, not automatic ground truth |

The table prevents a common category error: comparing an internet-enabled agent with a closed-corpus tool, or an analyst with paid databases against a system limited to public web pages. NIST recommends scaffolding that mirrors intended deployment (nvlpubs.nist.gov). It also recommends uniform execution-cost controls for a one-dimensional comparison (nvlpubs.nist.gov). If parity is impossible, publish performance and cost together rather than declaring a single winner.

# Build the Benchmark Pack

## Stratify the case set

Randomly selecting convenient companies will overstate performance. Build strata that force the failure modes acquisition research actually encounters:

- **Easy found facts:** clear legal name, location, and service description.
- **Aliases:** trading names, former names, abbreviations, and similar domains.
- **Acquired brands:** a brand site whose current legal owner differs from the visible brand.
- **Conflicting ownership:** authoritative sources with different dates or scopes.
- **Stale pages:** plausible but superseded statements.
- **Negatives:** evidence that explicitly rules out an answer.
- **No-answer cases:** a documented search yields no sufficient evidence.
- **Thin-footprint companies:** few indexable pages and limited registry coverage.
- **Changed facts:** ownership, leadership, location, or status changed around the cutoff.

Official systems illustrate why these strata matter. SEC search warns that a name may be listed differently than expected (www.sec.gov), while submissions data includes current and former names, exchanges, and tickers (www.sec.gov). Companies House exposes previous and dissolved-name search (www.gov.uk), and GLEIF supports fuzzy matching of names and addresses (www.gleif.org). These are useful signals, not universal coverage.

Use fictional entities or public benchmark entities, never a client list. Confirm licenses before distributing source snapshots. The SQuAD 2.0 paper, for example, explicitly records a **CC BY-SA 4.0** license (nlp.stanford.edu). A public URL does not itself grant unrestricted redistribution rights.

## Freeze the environment

Record the following before any candidate runs:

- **Mandate version:** definitions, answer classes, disqualifiers, scoring weights.
- **Corpus:** full snapshot, retrieval index, and cutoff date, or a documented live-web policy.
- **Affordances:** browsing, paid databases, code execution, file access, and memory.
- **System:** provider, model and version, temperature, prompt, tool versions, schema.
- **Budgets:** wall time, retries, tokens, queries, and source-access spending.
- **Locale:** jurisdiction, language, geography, and date conventions.
- **Split:** visible development set and hidden final test set.

Comparable evaluations should state whether systems have internet access (www.nist.gov). NIST also says benchmark versions should be associated with results (nvlpubs.nist.gov). Keep a hidden set because a public answer key invites adaptation to the test rather than the intended task. A representative transcript subset can support transparency while preserving that held-out set (www.nist.gov).

## Create gold answers and allowed unknowns

Two reviewers should independently assign a typed status and value, then reconcile disagreements without seeing candidate identities. The adjudication sheet should permit:

- **Supported:** one or more acceptable values with qualifying context.
- **Not found:** defined sources were checked and no sufficient evidence was located.
- **Conflicting:** credible sources disagree and the benchmark does not force a false resolution.
- **Not applicable:** the question does not apply under a written rule.
- **Unresolved:** the gold team cannot determine the answer reliably.

Registry data should not be treated as automatically verified truth. Companies House states that placement on the public record does not mean it verified the information (resources.companieshouse.gov.uk). GLEIF Level 2 data answers who owns whom, but specifically concerns accounting-consolidating parents (www.gleif.org). That scope is not identical to every mandate's ownership taxonomy.

*Table 2* is the minimum downloadable manifest schema.

| Field | Required content | Purpose |
|---|---|---|
| **case_id** | Stable non-semantic identifier | Joins runs without leaking the answer |
| **company_archetype** | Easy, alias, acquired brand, conflicting, stale, negative, no-answer, thin footprint, changed fact | Supports stratified error analysis |
| **question** | Exact mandate wording plus field definition | Fixes the construct being tested |
| **gold_status / gold_value** | Supported, not found, conflicting, not applicable, unresolved; normalized value if available | Prevents forced answers |
| **acceptable_variants** | Allowed aliases, units, date tolerance, categories | Makes scoring deterministic |
| **source record** | URL, excerpt, publisher, source date, access date, retained-copy key | Enables support and temporal review |
| **difficulty / risk** | Predeclared stratum and low, medium, or high consequence | Enables release gates by field |
| **split** | Development or hidden test | Limits test-set adaptation |

The manifest makes a benchmark inspectable. It also prevents corpus and label drift: TREC cautions that a document collection and its relevance judgments must match (trec.nist.gov). A changed corpus creates a new benchmark version, even if the questions are unchanged.

# Implementation Considerations and Process Changes

## Run repeated, logged trials

Run every case across every system, and run each system multiple times. Three runs are a useful minimum for detecting visible instability, not a claim of statistical adequacy. Increase repeats when outputs are variable, the field is high-risk, or the difference between contenders is small.

Each run log should retain:

- **Configuration:** system name, model version, prompt hash, toolset, corpus version.
- **Timing:** start, end, retrieval time, generation time, and review time.
- **Inputs and outputs:** question, answer, abstention, confidence if supplied.
- **Evidence:** retrieved URLs, excerpts, dates, and citation-to-claim mapping.
- **Trace:** tool calls, retries, intermediate decisions, and termination reason.
- **Economics:** input and output tokens, API charges, paid-source fees, analyst minutes.

NIST recommends saving full logs beside summary statistics (nvlpubs.nist.gov). OpenTelemetry's current generative-AI conventions include model names, token counts, and durations by default (opentelemetry.io). They exclude prompt content and tool arguments by default (opentelemetry.io), so a benchmark needs an explicit, access-controlled policy if those artifacts are required.

## Inspect traces, not scores alone

A passing final answer can still arise from an invalid path, such as seeing an answer key, using a prohibited source, resolving the wrong entity and later correcting by chance, or ignoring the time cutoff. NIST defines evaluation gaming by whether the method subverts measurement validity (www.nist.gov). It recommends closing task-design loopholes and setting clear rules (www.nist.gov).

Inspect all high-risk errors and a random sample of passes. Automated trace review can prioritize cases, but NIST expects it to remain paired with manual inspection (www.nist.gov). Strip sensitive or licensed content before wider transcript release; the draft guidance explicitly notes protection of sensitive transcript information (nvlpubs.nist.gov).

## Control privacy and licensing

Use fictional or public benchmark entities and synthetic mandates. Do not place client targets, private notes, or personal contact data into a reusable test set. The UK Information Commissioner's Office states the data-minimization principle plainly: process only personal information needed for the specified purpose (ico.org.uk). Document rights for every frozen page, registry extract, benchmark label, and redistributed artifact. If rights are unclear, retain a locator and a permitted short excerpt rather than packaging the full source.

## Operational checks before the hidden run

The benchmark owner should complete a final configuration review:

- **Web preservation:** account for the fact that information objects and web resources change over time (www.w3.org).
- **Evidence support:** distinguish citation presence from actual support.
- **Adaptation parity:** control the strategy used to adapt each language model to the scenario (crfm.stanford.edu).
- **Changed-fact coverage:** retain both historical and current LEI records when available (www.gleif.org).
- **Benchmark reproducibility:** record every condition needed to reproduce the run (mlcommons.org).
- **Registry interpretation:** remember that Companies House performs only basic checks when examining accounts (www.gov.uk).
- **Row-level results:** retain inputs, responses, explanations, and metric results for each evaluated row (docs.cloud.google.com).
- **Evaluator version:** record the grader model version for reproducibility (learn.microsoft.com).
- **Trace sampling:** use transcript review to identify task, tool, and integration issues (www.nist.gov).
- **Third-party inputs:** document treatment of third-party data and software (airc.nist.gov).
- **Cost instrumentation:** use token and duration metrics to estimate per-request cost (opentelemetry.io).
- **Prompt-change testing:** compare evaluation results after prompt changes (aws.amazon.com).
- **Continuous improvement:** maintain and continually improve the AI management system (committee.iso.org).

# Scoring and Human Adjudication

## Score claims and questions separately

Use exact-match or allowed-set scoring for typed answers, then a separate rubric for evidence. Three core formulas should remain distinct:

- **Claim-level precision** = supported correct claims / all claims made.
- **Coverage** = questions answered acceptably / questions required.
- **Abstention precision** = justified abstentions / all abstentions.

Add **citation faithfulness rate**, **identity accuracy**, **temporal-validity rate**, and **prohibited-inference rate**. Do not combine these into a single index until every non-compensable release gate has been evaluated. A severe identity error must not be canceled by high coverage elsewhere.

Source authority must be question-specific. SEC's public APIs require no authentication or API key (www.sec.gov) and update throughout the day as submissions are disseminated (www.sec.gov). By contrast, the SEC also notes that not every company offering stock must file electronically (www.sec.gov). Missing SEC evidence is therefore not proof that a private company lacks a feature or relationship.

## Use blinded human adjudication

Two reviewers should score independently, blinded to candidate identity where practical. Measure agreement, then send disagreements to a senior adjudicator with access to the frozen evidence, rubric, and allowed-answer set. NIST identifies multiple judges and inter-rater agreement as useful practices (nvlpubs.nist.gov).

Google likewise recommends evaluating model-based graders against human ratings used as ground truth (docs.cloud.google.com).

Machine graders can flag likely citation mismatches, missing qualifications, or inconsistent schemas. They should not settle disputed ownership, a disqualifying mandate fact, or the truth of an unresolved source conflict. Docket's published evidence model similarly keeps unresolved questions visible and uses a separate review pass. This is a process claim about Docket's documented design, not independent validation of its performance.

### Apply a fixed error taxonomy

Every rejected answer should receive one primary error code and optional contributing codes:

- **Wrong entity:** evidence belongs to another business.
- **Stale fact:** the source predates a relevant change or cutoff.
- **Unsupported inference:** the conclusion exceeds the cited material.
- **Citation mismatch:** the cited span discusses another claim.
- **Partial support:** only part of a compound answer is supported.
- **Missed answer:** sufficient evidence existed in the allowed corpus.
- **False answer instead of abstention:** the system should have returned unknown or conflict.
- **Source-quality failure:** the source class does not meet the field's evidence rule.

Analyze error concentration by question, company archetype, source type, and run. A low overall error rate can still conceal a release-blocking cluster in ownership or identity.

## Data Analysis and Evidence

### Estimate rates with uncertainty

For each proportion, publish the numerator, denominator, point estimate, and **95% Wilson confidence interval**. NIST's statistical handbook explains that the Wilson interval is based on inverting the hypothesis test (www.itl.nist.gov). Wilson intervals behave better than the simple normal approximation when samples are small or rates are near zero or one.

Compare contenders on paired cases because case difficulty otherwise confounds results. Report the paired difference with an appropriate standard error, a practice identified in NIST AI 800-2 (nvlpubs.nist.gov). Report effect size and interval, not only a significance label; NIST says significance should be interpreted probabilistically and alongside effect size (nvlpubs.nist.gov). If the interval spans both a meaningful win and loss, the result is inconclusive.

A tiny benchmark cannot establish superiority. For illustration, a system with 19 acceptable answers from 20 cases has a 95% Wilson interval of approximately **76.4% to 99.1%**, despite a 95% point estimate. The calculation is illustrative, not a measured vendor result. Expand the sample in the strata that drive the decision rather than adding many low-risk easy cases.

### Measure economics at the accepted-answer level

Use:

**Cost per accepted answer = (run cost + source cost + reviewer cost) / accepted answers.**

Report p50 and p90 elapsed time, analyst review minutes, retries, and acceptance rate. Prometheus defines the 0.5 quantile as the median (prometheus.io) and identifies 0.95 as the 95th percentile (prometheus.io). For this benchmark, p90 is often more operationally useful than an average because a research queue is delayed by slow-tail cases.

As a current price anchor, Amazon Bedrock lists **$0.21 per completed human evaluation task**, excluding model inference (aws.amazon.com). That figure is not a universal reviewer-cost estimate. Teams should substitute their own loaded analyst rate and measured minutes. The relevant comparison is not cheapest output, but cheapest answer that survives the acceptance gate.

### Set release gates by field

*Table 3* provides a procurement release matrix. Thresholds are deliberately expressed as organization-set requirements rather than invented universal values.

| Field | Primary metric | Required gate | Human review |
|---|---|---|---|
| **Identity** | Entity-resolution precision | Set an extremely high minimum; any wrong-entity cluster blocks release | Mandatory for ambiguous aliases, shared domains, and acquired brands |
| **Ownership** | Correctness, authority, temporal validity | No unsupported ownership inference; fresh dated source required | Mandatory before outreach or mandate exclusion |
| **Mandate fit** | Per-question precision and coverage | Each disqualifier passes independently; no averaging across fields | Mandatory for consequential exclusions |
| **Revenue estimate** | Method compliance and interval disclosure | Must remain labeled estimate; unknown allowed; no promotion to verified fact | Mandatory if used in ranking |
| **Contact** | Identity match, freshness, permitted-source compliance | Exact person-company match and collection date required | Sample review plus review of ambiguous matches |
| **Unknown handling** | Abstention precision and not-found documentation | Justified unknowns accepted; false certainty blocks release | Review all high-risk abstentions and conflicts |

The matrix turns an aggregate leaderboard into a deployment decision. A candidate may pass identity and fit while remaining unsuitable for ownership or revenue. That outcome supports a limited release by question type, which is more informative than a blanket pass or fail.

## Procurement Decision and Release Process

Use the benchmark in five stages:

1. **Pre-register the decision.** State use cases, non-negotiable gates, tie rules, budget, and who may adjudicate.

2. **Run the visible development set.** Allow candidates to fix schemas and obvious integration problems without revealing hidden answers.
3. **Freeze configurations.** Hash prompts, version tools, record paid-source access, and prohibit candidate-specific post-processing.
4. **Run the hidden set.** Randomize order, retain complete traces, and blind reviewers to candidate identity where possible.
5. **Issue a field-level decision.** Pass, limited deployment, remediate and retest, or reject for each question type.

Do not publish vendor scores without the full case specification, source-access parity, sample size, intervals, and adjudication method. NIST recommends exact model versions and protocol details in reports (nvlpubs.nist.gov). Item-level results are particularly useful for comparing evaluator runs (nvlpubs.nist.gov). Redact protected information, but retain enough structure for an internal audit.

Release should be conditional on monitoring. NIST says AI systems should be tested before deployment and regularly during operation (airc.nist.gov). Re-run the hidden monitoring set after any change to the model, prompt, tool, retrieval corpus, source permissions, schema, or scoring rule. NIST's playbook also calls for procedures that track dataset modifications (airc.nist.gov). Treat a material change as a new system version.

## Implications and Future Directions

The practical consequence is a shift from vendor selection to **use-case authorization**. A team can approve a system for identity checks and descriptive fit while requiring human confirmation for ownership. It can permit automated not-found triage while prohibiting a negative conclusion from absence alone. This is safer and more economically legible than demanding one universal accuracy score.

Evidence infrastructure will matter as much as model choice. NIST's evaluation-probes project aims to produce structured audit trails mapping decisions to documents (www.nist.gov). W3C defines provenance information as useful for assessing quality and reliability (www.w3.org). For acquisition research, that means the durable asset is not only the answer. It is the answer, source state, excerpt, run, reviewer decision, and supersession history.

Future benchmarks should expand carefully. Add jurisdictions only after modeling their registries and ownership concepts. GLEIF says its Registration Authorities List spans **232 jurisdictions** and more than **1,050** business registers and related authorities (www.gleif.org). Breadth of mapping does not make legal concepts uniform. Maintain jurisdiction-specific rules and preserve conflicts instead of forcing a global label.

## Frequently Asked Questions (FAQs)

### How should a team benchmark an AI research agent?

Define the intended decision, score company-question-date cases, freeze sources and affordances, create two-reviewer gold answers, run every candidate repeatedly, retain traces, and apply field-specific release gates. This is the core **AI agent evaluation framework for private equity**. It evaluates the configured workflow, not just a named model.

### What are the most useful AI research agent accuracy metrics?

Use claim-level precision, coverage, abstention precision, citation faithfulness, completeness, sufficiency, identity accuracy, temporal validity, and prohibited-inference rate. Report each separately with its numerator, denominator, and confidence interval. MLCommons similarly identifies reproducibility as a core benchmark objective (mlcommons.org).

### How can a PE team evaluate AI deal-sourcing tools fairly?

Give systems the same mandate, cases, cutoff, source access, geography, language, time limit, and acceptance rules. If an agent has the web, a platform has proprietary data, and an analyst has paid sources, report those differences and analyze performance with cost. Do not interpret throughput as a deal, conversion, or origination outcome.

### What belongs in acquisition-target research benchmarks?

Include easy facts, aliases, acquired brands, conflicting ownership, stale sources, explicit negatives, justified no-answer cases, thin-footprint companies, and changed facts. Weight the set toward the questions that matter to the mandate, then keep a hidden test split.

### Can automated graders replace human review for private-equity AI due diligence tools?

No. Automated graders can triage citation and format problems, but high-risk ownership, identity, and disqualifying facts need fresh human review. NIST says independent review can improve testing effectiveness (airc.nist.gov).

### How should benchmarking AI agents for company research handle unknowns?

Treat supported not-found, conflicting, not applicable, and unresolved as allowed gold states. Measure abstention precision separately from coverage. A justified unknown is a useful research result, while a fabricated categorical answer is a failure.

### Which evidence quality metrics matter most for AI research agents?

Citation faithfulness tests whether the source supports the claim, completeness tests whether qualifications were preserved, sufficiency tests whether the source is strong enough, and authority tests whether the source class fits the field. Temporal validity and entity match must be checked separately.

## Conclusion

An AI research agent benchmark for acquisition-target research is an evidence-control system, not a writing contest. Its unit is the company-question-date case. Its foundation is a stratified, licensed corpus with explicit unknown states, two-reviewer gold answers, frozen configurations, repeated runs, and complete traces. Its outputs are separate measures for correctness, evidence quality, coverage, abstention, identity, time validity, and inference discipline.

The procurement decision should be equally specific. Pass or limit systems by question type, require fresh human review for ownership and consequential exclusions, and calculate cost per accepted answer rather than cost per generated record. Use Wilson intervals and paired case-level comparisons to show uncertainty, and state when the sample is too small to distinguish contenders.

Finally, re-test after every material model, prompt, tool, corpus, schema, or policy change. A well-designed benchmark does not promise deal success. It establishes which research questions a configured workflow can answer to a documented standard, where it must abstain, how much review it consumes, and what evidence a decision maker can inspect.

## Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.


1. https://www.w3.org/TR/prov-constraints
2. https://www.gov.uk/guidance/searching-the-companies-house-register
3. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-2.ipd.pdf
4. https://learn.microsoft.com/en-us/azure/ai-services/speech-service/how-to-voice-live-evaluate
5. https://www.nist.gov/programs-projects/building-evaluation-probes-agentic-ai
6. https://github.com/usnistgov/agentic-research-evaluation-probes
7. https://www.nist.gov/caisi/guidelines
8. https://www.rfc-editor.org/rfc/rfc7089.html
9. https://www.sec.gov/search-filings/cik-lookup
10. https://www.sec.gov/search-filings/edgar-application-programming-interfaces
11. https://www.gleif.org/en/lei-data/access-and-use-lei-data
12. https://nlp.stanford.edu/pubs/rajpurkar2018squad.pdf
13. https://www.nist.gov/caisi/cheating-ai-agent-evaluations/4-practices-detecting-and-preventing-evaluation-cheating
14. https://resources.companieshouse.gov.uk/disclaimer.shtml
15. https://trec.nist.gov/data/reljudge_eng.html
16. https://opentelemetry.io/blog/2026/genai-observability
17. https://www.nist.gov/caisi/cheating-ai-agent-evaluations
18. https://ico.org.uk/about-the-ico/research-reports-impact-and-evaluation/research-and-reports/technology-and-innovation/tech-horizons-and-ico-tech-futures/ico-tech-futures-agentic-ai/data-protection-and-privacy-risks
19. https://crfm.stanford.edu/2022/11/17/helm.html
20. https://mlcommons.org/benchmarks
21. https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/eval-python-sdk/view-evaluation
22. https://airc.nist.gov/airmf-resources/airmf/5-sec-core
23. https://aws.amazon.com/bedrock/evaluations
24. https://committee.iso.org/cms/live/live/en/sites/isoorg/contents/news/insights/AI/what-is-ai-all-you-need-to-know/newsBody/standard-reference/standard-reference%40/81230.html

25. https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/evaluate-judge-model
26. https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm
27. https://prometheus.io/docs/practices/histograms
28. https://aws.amazon.com/bedrock/pricing?nc2=h_pr_br
29. https://airc.nist.gov/airmf-resources/playbook/manage
30. https://www.w3.org/TR/prov-overview
31. https://www.gleif.org/en/lei-data/code-lists/gleif-registration-authorities-list

THE PUBLISHER

# About Docket

Docket provides deal-origination research software and managed research for private equity firms. We help investment teams investigate acquisition targets using structured screening criteria, retained sources and reviewable company evidence. Teams can work through a self-serve platform or use managed research, depending on how they want research delivered.

### Research against a defined mandate

A useful target list needs more than company names. Docket focuses on the questions that determine whether a company fits an investment mandate, including the evidence needed to support or qualify each answer. Our research approach makes the connection between screening criteria, source material and conclusions visible to the team reviewing the work.

### Triage, Scout and Audit

Docket's three named research agents perform complementary tasks. Triage resolves company identity and screens fit. Scout collects sourced answers against the mandate. Audit checks retained evidence, addresses contradictions and leaves unsupported answers visibly unresolved. This structure helps reviewers distinguish established findings from missing information and questions requiring further investigation.

### Evidence that supports investment-team judgment

Our research library covers market mapping, screening criteria, private-company data, succession and ownership, source evaluation and evidence standards. These resources explain the methods and limitations behind origination research. Findings support a team's judgment; they do not establish that a company is for sale or guarantee a transaction or investment outcome.

### Work with Docket

Visit Docket to explore the platform, managed research and the current contact path. Read about screening criteria, evidence standards, AI research agents and market mapping.

Public examples are illustrative unless explicitly identified otherwise. Research preparation and authorized outreach are separate activities; confidential target lists and customer outcomes should never be inferred from an educational example.

**Website:** https://docket.capital

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.