Docket

DOCKET / RESEARCH & PRACTICAL GUIDANCE

Private Equity Research Agent Source Access Test (2026)

> 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 · 2026-11-15 · private equity research agent · source access test · ai agent acceptance test · access-state ledger · false closure rate · deal sourcing ai evaluation · robots.txt ai agents · acquisition target research · ai due diligence · vendor evaluation criteria

Original article: https://docket.capital/articles/source-access-acceptance-test-pe-research-agents



In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes to How Missing Answers Are Judged
4. Controlled Acceptance Pack and Paired Review
5. Implementation Considerations and Process Changes
6. Data Analysis and Evidence
7. Case Studies and Real-World Examples
8. Implications and Future Directions
9. Frequently Asked Questions (FAQs)
10. Conclusion

Executive Summary

A private equity (PE) research agent that answers "ownership not found" or "no mandate fit evidence" has said one of two very different things: the evidence does not exist, or the agent never read the page that holds it. This report proposes an acceptance test that separates the two. It is a per-question disposition and an access coverage denominator, not a new overall agent benchmark, and it is not a guide to bypassing robots rules, paywalls, logins or bot protection.

The core idea is an **access-state ledger** with five stages per question: source discovery, permission assessment, retrieval, render and extraction, and evidentiary sufficiency. A cited URL or a successful HTTP request proves none of the later stages. One July 2026 controlled study of **10 assistants** and **200 trials** compared server logs with answer correctness and concluded the two "must be analyzed separately" (arxiv.org). One vendor's tool documentation makes the same point: fetch errors are returned inside a **200 (success)** response body (platform.claude.com).

The report defines a pack of **8 fictional page types** (readable, disallowed path, challenge page, authorization wall, empty app shell, scanned document, transient error, absent fact), a source-attempt schema, and two headline metrics. **Access coverage** is relevant permitted sources whose content was actually obtained divided by relevant permitted sources attempted. **False closure rate** is questions marked no-evidence despite a qualifying fact on an approved test path, divided by all fixture questions marked no-evidence. Both come with Wilson intervals, which are recommended for small samples (projecteuclid.org) (www.itl.nist.gov).

The environment justifies the test. Among **16,213,084** sites in the 2025 Web Almanac crawl, about 85% returned a valid robots.txt, while roughly 13% returned 404, 1.0% timed out and 0.5% returned 403 (almanac.httparchive.org). A challenge page is an HTML response that interrupts an expected API or XHR (XMLHttpRequest) flow (developers.cloudflare.com). A practitioner audit dated 2026-07-16 calls private-markets intelligence a near dead zone for AI retrieval, an attributed opinion rather than a measured rate (davidkemp.ai).

Decision rule: accept a batch only when every question carries a disposition (accepted evidence, retry, approved alternative route, analyst review, unresolved after a read source, or out of scope), when coverage is reported by source class and question risk, and when unresolved mandate disqualifiers stay on hold regardless of aggregate accuracy. The primary study is a short-window preprint on simple pages, not a measured error rate for any PE workflow (arxiv.org).

Introduction and Background

PE origination and diligence teams ask research agents narrow questions: who owns this company, does it provide the service the mandate targets, is there a disqualifying fact. The hardest outcome to audit is the empty one. An answer of "no ownership information found" is accepted by many review processes as a legitimate result, yet it can arise from at least two upstream causes that look identical in the final text. This report examines how to tell them apart before a batch is accepted.

Adoption context matters for urgency. A Bain and StepStone survey of **103** investment and investor relations professionals, fielded December 2025 to January 2026, found general partners most often pointing to due diligence and deal sourcing as the highest-return generative AI use cases (www.bain.com). A Deloitte survey of 1,000 US corporate and PE leaders reported that **86%** of responding organizations had integrated generative AI into merger and acquisition workflows, and 67% named data security as a leading concern, followed by data quality and availability (blended respondents, first half of 2025) (www.deloitte.com). Neither survey measures source-access failure.

The analytical contribution here is narrow. Existing agent benchmarks score answers and citations. BrowseComp, for example, comprises **1,266** questions (arxiv.org), and Mind2Web 2 has **130** long-horizon browsing tasks (arxiv.org). These ask whether the final answer is right. This test asks an upstream question: did the agent obtain the relevant passage, and if not, which stage failed and who decides what happens next. Docket, the publisher of this report, offers sourced target research through Triage, Scout and Audit agents; the acceptance test below is written so that it can be applied to any vendor's agent, including the publisher's, and the report makes no performance claim for any system.

Three terms recur. A **fixture** is a locally controlled or expressly permitted test page with a known expected access status. A **qualifying fact** is a passage on an approved path that would answer the question under the mandate. An **approved path** is a route the organization is entitled to use: an official API, a licensed subscription, a registry download, or an analyst working under the source's own terms. The ISTQB (International Software Testing Qualifications Board) glossary defines acceptance testing as a test level focused on determining whether to accept the system (glossary.istqb.org), which is the decision this report supports.

Scope boundaries are deliberate. RFC 9309, the Robots Exclusion Protocol standard, states that its rules "are not a form of access authorization" (www.rfc-editor.org), and Cloudflare documents challenge pages as a control that site owners configure (developers.cloudflare.com). Nothing below recommends circumvention: a policy-restricted source is a routing decision, not a technical defect to engineer around.

Key Changes to How Missing Answers Are Judged

Change 1: Define the claim and the required source classes first

Counting attempts before defining what counts as evidence produces flattering denominators. The first change is to write, per question, the claim being tested and the source classes that could settle it. A registry entry can settle legal identity; a consolidated-accounts parent link can suggest control; a product page or filing can speak to service fit. One class rarely substitutes for another.

Jurisdiction decides what exists to be found. The US Financial Crimes Enforcement Network (FinCEN) states that domestic companies are exempt from federal beneficial ownership information reporting under its 2026 final rule (www.fincen.gov), and that reported information sits in a non-public database (www.fincen.gov). A "no federal beneficial owner record" result for a US private company is therefore the expected state, not evidence about that company. In the UK, a person with significant control (PSC) is someone who owns or controls a company (www.gov.uk), and Companies House publishes a free PSC snapshot (download.companieshouse.gov.uk). A UK question has an approved registry path; the same question about a US private company mostly does not.

The question record should therefore carry a **mandate version** and required source classes. When a source class is out of reach for legitimate reasons, the correct disposition is "out of scope for this route", not "no evidence".

Change 2: Record a five-stage access pipeline, not a citation list

The ledger tracks each question through five stages, and each stage has its own evidence.

- **Discovery:** the URL or document was identified, and by what means. A snippet found in search is a lead, not evidence.
- **Permission assessment:** the applicable rules and the permitted interface were checked and recorded as a snapshot. RFC 9309 notes that rules apply per origin and match user agents case-insensitively (www.rfc-editor.org), and Google documents that rules apply only to the host, protocol and port where the file sits (developers.google.com).
- **Retrieval:** timestamp, tool, status, redirect chain, final URL and content type.
- **Render and extraction:** whether the body contained the substantive content, and how it was converted to text.
- **Evidentiary sufficiency:** the excerpt identifier that supports or contradicts the claim, or an explicit failure category.

The separation of stages mirrors the structure of the July 2026 preprint "Do Generative AI Assistants Respect robots.txt?". Its authors used controlled domains and server-side logs, and placed a secret code (an HMAC, hash-based message authentication code, over the page version) on each target page so that retrieval could be verified separately from the answer (arxiv.org). A configuration counted as working only when both the log and the answer agreed (arxiv.org). The authors state that server-side retrieval and user-visible answer correctness "are related but not equivalent" (arxiv.org). A PE team cannot plant secret codes on real registries, but it can adopt the principle: require an independent trace that the relevant page was obtained. Black-box

tools limit this, since retrieval cannot be traced to the component that fetched the page (arxiv.org). Where a vendor exposes only final answers, empty ledger cells should be reported as unobservable rather than filled by inference.

Change 3: Use a failure taxonomy that keeps not-retrieved apart from read-but-absent

The central distinction is between states where the agent never obtained usable content and states where it read the source and the field was not there. *Table 1* summarizes the taxonomy, with the typical signal, what the state does and does not prove, and the default disposition.

State	Typical signal	What it proves	Default disposition
Not searched	No discovery record for a required source class	Nothing about the evidence	Retry within scope, or analyst review
Policy-restricted	Rule snapshot disallows the path for the agent's identity	Access not permitted for this interface, not that the fact is absent	Approved alternative route, or out of scope
Authorization required	HTTP 401 or a login form (www.rfc-editor.org)	Credentials needed; content unread	Licensed or API route, or analyst under source terms
Forbidden or hidden	HTTP 403 or 404; a server may use 404 to avoid disclosing a resource (www.rfc-editor.org)	Refusal or non-disclosure, not proof of nonexistence	Analyst review
Rate limit	HTTP 429 (www.rfc-editor.org)	Too many requests in a window	Time-boxed retry with back-off
Transient or timeout	HTTP 503 temporary overload (www.rfc-editor.org), network timeout	Service temporarily unavailable	Time-boxed retry, then escalate
Interstitial or challenge	HTML challenge body where data was expected (developers.cloudflare.com)	A challenge was read, not the target	Approved API route, or analyst review
Dynamic content not rendered	200 status, empty application shell	Shell retrieved, content unread	Permitted rendering path or API
Extraction failure	Scanned or image-only document, unsupported type (platform.claude.com)	Document obtained, text not extracted	Permitted OCR or analyst reading
Read, field absent	Substantive body parsed, no qualifying passage	Absence within this source only	Unresolved; check other source classes
Read, contradicts	Passage conflicts with claim	Evidence against the claim	Record contradiction; analyst review

Only the last two rows are read states. Everything above them is a retrieval or extraction state, and none of them should be reported as "no evidence". The status codes show why: RFC 9110 explains that a 404 does not indicate whether the absence is temporary or permanent and that servers may prefer not to disclose a resource (www.rfc-editor.org), and RFC 7725 notes that a 451 response (unavailable for legal reasons) implies neither the existence nor the nonexistence of the resource (www.rfc-editor.org).

The table is a working taxonomy for audit, not a protocol standard. Two states deserve extra care. A challenge page arrives with a different body than expected; Cloudflare documents a header on challenge responses that lets a client detect them (developers.cloudflare.com), and its error-page reference lists the status codes such pages use (developers.cloudflare.com). An agent that stores the interstitial text as "the page" has recorded a retrieval success and an evidentiary failure at once. For dynamic pages, a December 2024 crawler-level log analysis of OpenAI, Anthropic, Meta, ByteDance and Perplexity crawlers found that none of the major AI crawlers it examined render JavaScript, while reporting Google's Gemini (which uses Googlebot's infrastructure) and AppleBot as exceptions (vercel.com). That finding is dated and crawler behaviour changes. Whether a particular agent renders is a property to test, not assume.

Robots files themselves have separate handling rules that must not be confused with target-page handling. RFC 9309 says a 4xx on the robots.txt fetch means the file is "unavailable" and the crawler may access any resources (www.rfc-editor.org), while server or network errors make the file "undefined" and the crawler must assume complete disallow (www.rfc-editor.org). Google documents its own variants for 5xx (developers.google.com). These rules govern how a crawler treats the robots.txt file; they say nothing about whether a target page's content was read, and they do not override a site's terms.

Controlled Acceptance Pack and Paired Review

The fixture pack

The pack uses fictional companies on pages the testing team controls, or on sources that expressly permit testing; it never probes real companies' defenses. Each fixture has a manifest entry stating the expected access status, expected evidentiary status, controlled source version, and expected review action. *Table 2* lists eight fixture types.

Fixture	Controlled behavior	Expected access status	Expected evidentiary status	Expected review action
F1 Readable page	Static HTML states a fictional owner	Retrieved	Qualifying fact present	Accept evidence
F2 Disallowed path	Rule snapshot disallows the path for the test agent identity	Not retrieved (policy)	Fact exists on approved copy	Route to approved path
F3 Challenge response	Local server returns an HTML interstitial with a marker	Retrieved wrong body	Not read	Route or analyst; never "no evidence"
F4 Authorization wall	Returns 401 and a login form	Not retrieved (auth)	Fact exists on licensed copy	Licensed route
F5 Empty app shell	200 with a shell; content only via a script call	Retrieved, not rendered	Not read	Permitted render or API route
F6 Scanned document	Image-only PDF stating the owner	Retrieved, not extracted	Fact in image	Permitted OCR or analyst
F7 Transient error	503 with Retry-After, then 200	Recoverable	Qualifying fact after retry	Time-boxed retry

Fixture	Controlled behavior	Expected access status	Expected evidentiary status	Expected review action
F8 Absent fact	Readable page, field genuinely missing	Retrieved	No qualifying fact	Unresolved; keep as legitimate result

Table 2 is only useful if each fixture is validated before use. A challenge page or a snippet must not be able to pass as a source document, so the validation step checks the body for the planted marker text and the content type, not just the status code. An external tool's own error taxonomy shows how fine the states can be: Anthropic's web fetch tool documents a distinct list of error codes, including inaccessible URL, disallowed URL, rate limit and unsupported content type (platform.claude.com), and counts failed fetches against the tool's use limit (platform.claude.com). A harness that records only "tool returned" would collapse these states.

F8 protects against over-correction: without a genuinely absent fact, an agent could be rewarded for never saying "unresolved". The test needs both directions, failing to read when a fact exists and claiming a fact the page lacks. F2 tests restraint, and the preprint argues that avoiding a disallowed page is meaningful only when the same system demonstrates successful retrieval elsewhere (arxiv.org), so every negative control is paired with a positive one.

These failure modes are plausible, not exotic. A 2026 silent-failure audit injected empty or malformed payloads behind HTTP 200 responses and found fabrication, where the agent treats an empty payload as real data, in 56.6% of valid responses (arxiv.org). It used four models and twelve stubbed tools, so it illustrates mechanism rather than a rate for any PE workflow.

Paired review

The second control uses the same questions and cutoff date for two routes: the agent, and an authorized analyst or official API route. Reviewers adjudicate one thing per question: was the relevant passage obtained, and by which route. A manual path is not a gold answer if the underlying evidence is absent or ambiguous. In that case the pair agree on "unresolved" and the fixture record notes that no qualifying fact exists.

Agreement between reviewers should be measured. Interrater reliability is the extent to which raters assign the same score to the same variable (pmc.ncbi.nlm.nih.gov), and McHugh's worked example shows percent agreement of 0.94 alongside a Cohen's kappa of 0.85 (pmc.ncbi.nlm.nih.gov). Report both when two reviewers adjudicate the same entries.

Where a vendor gives no trace, the paired design still works at the answer level: compare the cited passages and mark entries "trace unavailable", and say so in the batch report.

From state to disposition

Each question ends in exactly one disposition, and time-boxing is part of the rule. The suggested set is:

- **Accepted evidence:** the relevant passage was obtained and the excerpt stored.
- **Retry:** a transient or rate-limit state, within a stated budget.
- **Approved alternative route:** a policy, authorization or challenge state with an official or licensed path.
- **Analyst review:** unrendered, unextracted or contradicting content.

- **Unresolved after read:** sources read across the required classes, field absent.
- **Out of scope:** the source class is unreachable for this route or jurisdiction.

A retry budget (for example, a fixed number of attempts over a fixed window) should respect documented limits: Companies House allows up to 600 requests in a five-minute window and returns 429 beyond it (developer-specs.company-information.service.gov.uk), and the SEC states a current maximum of 10 requests per second for EDGAR (Electronic Data Gathering, Analysis, and Retrieval) (www.sec.gov). Exceeding a documented limit turns a transient defect into a policy problem. Policy-blocked sources are never retried by technical means; they move to an approved route or to analyst review.

Implementation Considerations and Process Changes

Approved routes by evidence need

The alternative to "no evidence" is an approved route with its own coverage limits. *Table 3* summarizes routes relevant to identity and ownership questions, the limits that each documents, and how the acceptance ledger should treat them. It also includes agent-based routes themselves, because a vendor's agent is one more route that should emit the same ledger.

Route	Documented access terms	Coverage limit that affects the ledger	Ledger treatment
SEC EDGAR APIs	Public JSON APIs without authentication (www.sec.gov); automated tools must follow the fair-access policy (www.sec.gov)	Covers registrants required to file (www.sec.gov); not a register of private-company owners	Absence is "outside route coverage" for non-filers
UK Companies House API and PSC snapshot	API key (developer-specs.company-information.service.gov.uk); 600 requests per five minutes (developer-specs.company-information.service.gov.uk)	UK-registered companies; PSC snapshot is unsupported (download.companieshouse.gov.uk)	Jurisdiction recorded; snapshot date recorded
GLEIF Level 2	Free access, CC0 licence (www.gleif.org); 60 requests per minute per user (api.gleif.org)	Parent links exist only where the parent holds a Legal Entity Identifier (www.gleif.org); parent means accounting consolidation (www.gleif.org); reporting may be declined (www.gleif.org)	Absent parent is "not reported", not "no parent"
EU beneficial ownership registers	National register links via the e-Justice portal (e-justice.europa.eu)	Terms vary by Member State	Record country and entitlement
Commercial registry aggregator API	API key required (api.opencorporates.com); terms bar passing API data to third parties (opencorporates.com)	Licence terms limit onward use	Check licence against report use

Route	Documented access terms	Coverage limit that affects the ledger	Ledger treatment
Vendor research agents (including the publisher's own Triage, Scout and Audit)	Vendor-defined	No performance claim is made here; trace visibility varies	Same ledger and fixture pack apply

The table maps constraints and does not recommend any provider. Two reading rules follow. First, "absence" must carry the route's coverage statement, because several sources document structural gaps: GLEIF records only relationship data the entity is willing to report, and the US federal regime currently exempts domestic companies. Second, a route's terms determine whether its output can enter a client deliverable at all.

Distinguishing agent identities and policies

Training crawlers, search bots and user-triggered retrieval can have different policies. OpenAI documents separate robots for search, training and user actions and states that each setting is independent (developers.openai.com); it also says that for user-initiated actions robots.txt rules may not apply (developers.openai.com). Anthropic documents three separate robots for training, user-initiated retrieval and search (support.claude.com). Google separates automatic crawlers, which respect robots.txt (developers.google.com), from user-triggered fetchers, which generally ignore it (developers.google.com), and Perplexity documents a comparable user-action fetcher (docs.perplexity.ai). These are documentation statements as of November 2026.

The consequence is procedural: record which identity the agent presented, because a policy decision made for one identity does not carry to another. Generic user agents make attribution from server logs difficult (arxiv.org). Cloudflare describes verified bots as transparent about who they are and what they do (developers.cloudflare.com) and documents Web Bot Auth, which uses cryptographic signatures in HTTP messages (developers.cloudflare.com). Whether a given agent supports such identification is a vendor-evaluation question. Anthropic states that it will not attempt to bypass CAPTCHAs for sites it crawls (support.claude.com), so a challenge page is a legitimate stop state to be routed, not defeated.

Governance expectations for vendor evaluation

The Financial Industry Regulatory Authority (FINRA) says member firms should evaluate generative AI tools before deploying them (www.finra.org), and its 2026 oversight report notes that multi-step agent reasoning can make outcomes difficult to trace, complicating auditability (www.finra.org). The International Organization of Securities Commissions (IOSCO) toolkit says documentation and logs should be maintained throughout the AI lifecycle (www.iosco.org) and that compliance and risk functions should be able to conduct due diligence on any third-party provider (www.iosco.org). The UK Financial Conduct Authority states that it does not plan extra AI-specific regulation and relies on existing frameworks (www.fca.org.uk). The Securities and Exchange Commission (SEC) fiscal 2026 examination priorities mention reviewing the accuracy of AI-related representations (www.sec.gov). These statements carry different legal force and are context, not a compliance checklist; readers should consult their own compliance advisers.

The NIST AI Risk Management Framework describes test, evaluation, verification and validation (TEVV) as a regular lifecycle activity and calls for documenting test sets, metrics and tools (nvlpubs.nist.gov); its Generative AI Profile adds knowledge limits and upstream data sources (nvlpubs.nist.gov). A fixture manifest and source-attempt ledger are instances of that documentation.

The source-attempt schema

Procurement questions for any vendor follow from the ledger:

- **Identity:** which user agent and tool identity does each fetch present?
- **Per-source trace:** are final URL, status and content type returned for every source?
- **State separation:** are challenge, authorization and rate-limit states distinguished from not-found?
- **Rendering and extraction:** does the tool execute scripts and read scanned documents?
- **Audit export:** can raw retrieved content be exported for review?
- **Retry accounting:** how are retries counted and limited?

The schema fields are: company or internal ID; question ID; mandate version; source URL and class; user agent, tool and interface; applicable rule snapshot or reference; attempt timestamp; response status; final URL; content type; extraction outcome; excerpt ID; failure category; approved fallback; reviewer; and final disposition. Store the rule snapshot as a hash and a date: RFC 9309 allows caching for up to about 24 hours (www.rfc-editor.org), and OpenAI notes its systems may take about 24 hours to adjust after an update (developers.openai.com), so the effective rule at attempt time may differ from today's. Store the extracted excerpt, not only a pointer, so a reviewer can check the passage without re-fetching a page that may have changed.

Data Analysis and Evidence

Metric definitions and denominators

Two measures carry the decision. **Access coverage** equals relevant sources whose content was actually obtained, divided by relevant permitted sources attempted. The denominator excludes policy-restricted sources (they move to an approved route) and counts only sources inside the permitted interface. It must never be labelled whole-web recall, because it does not estimate what exists on the web. **False closure rate** equals questions classified no-evidence despite an approved test path containing a qualifying fact, divided by all fixture questions marked no-evidence (falsely or correctly). This is not sensitivity, which is the proportion of true positive tests out of all patients with a condition (www.ncbi.nlm.nih.gov). The base here is the set of no-evidence calls, so the metric is the share of negative calls that are wrong, the complement of negative predictive value. It therefore depends on how many F8 absent-fact items the pack contains, and because the fixture items are clustered by company, the Wilson intervals below assume independence and are approximate.

A batch report should contain:

- **Coverage by source class,** with numerators and denominators.
- **Coverage by question risk tier.**

- **False closure rate,** with its interval.
- **Access-blocked share** and **error-to-retry recovery.**
- **Mandate disqualifiers** that remain unresolved, listed as holds.

Two supporting measures are the **access-blocked share** (questions whose first-attempt state is policy, authorization or challenge, over all questions) and **error-to-retry recovery** (transient-error fixtures that succeed within the retry budget, over all transient-error fixtures). AbstentionBench evaluates abstention across 20 datasets (arxiv.org) and measures the proportion of responses where the model correctly abstained (arxiv.org), and SQuAD 2.0 established the unanswerable-question format with over 50,000 items (arxiv.org). Those measure whether a model abstains appropriately given its information; the fixture pack adds whether the information was reachable in the first place.

Worked arithmetic (Hypothetical Example)

The following numbers are invented to show the calculation. They are not results for any real agent. A pack has 8 fixture types with 5 fictional companies each, for 40 questions.

- **Questions closed as no-evidence by the agent:** 18, of which 5 are F8 genuine absences (correct) and 13 have a qualifying fact on an approved path (false closures). False closure rate is 13 of 18, or 72.2%, with a 95% Wilson interval of about 49.1% to 87.5%.
- **Access coverage:** permitted sources attempted are F1, F3, F5, F6, F7 and F8, which is 30 sources. Suppose relevant content was obtained for 17 (F1 5, F3 0, F5 1, F6 2, F7 4, F8 5). Coverage is 17 of 30, or 56.7%, with an interval of about 39.2% to 72.6%.
- **Access-blocked share:** F2, F3 and F4 are 15 of 40 questions, or 37.5% (interval about 24.2% to 53.0%).
- **Error-to-retry recovery:** 4 of 5 transient errors recovered, or 80.0%, with an interval of about 37.6% to 96.4%. Five trials cannot support a tight claim, and the interval says so.

Wilson is the appropriate default: Brown, Cai and DasGupta recommend the Wilson or equal-tailed Jeffreys interval for small samples (projecteuclid.org), and the NIST handbook advises an exact binomial interval when the failure count or sample size is very small (www.itl.nist.gov). If an agent produced zero false closures among 18 no-evidence calls, the rule of three would put the 95% upper bound near 3/18, about 16.7% (the exact binomial bound is closer to 15.3%), a reminder that "zero observed" is not "zero rate" (pmc.ncbi.nlm.nih.gov).

Environment base rates

Fixtures do not show how often each state occurs in the wild; published aggregates give context but are not PE retrieval rates.

- **Robots file status:** In the 2025 Web Almanac (HTTP Archive crawl of July 2025, **16,213,084** websites) about 85% of robots.txt requests returned 200 (almanac.httparchive.org). The chapter reports timeouts near 1.0%, 403 responses near 0.5% and 5xx near 0.1% (almanac.httparchive.org). The wildcard user agent appeared in 77% of files (almanac.httparchive.org).

- **AI crawler purpose:** Cloudflare reported that about 80% of AI crawling was for training, 18% for search and 2% for user actions over a 12-month period (blog.cloudflare.com). A later Cloudflare report states 52% of crawler requests were for AI training as of June 2026 under a different set of categories (blog.cloudflare.com); the two splits use different denominators and should not be joined into a trend.
- **Traffic and site controls:** Cloudflare reported that human traffic generated 47% of HTML requests and non-AI bots 44% through 2 December 2025 (blog.cloudflare.com), and that fewer than 1% of sites on its network block search crawlers while 17% block training (blog.cloudflare.com). These are network observations, not whole-web measures.
- **Paywall context:** the Reuters Institute reports that the proportion paying for any online news is stable at 18% across 20 richer countries (reutersinstitute.politics.ox.ac.uk), from an online survey fielded mid-January to end of February 2025 (reutersinstitute.politics.ox.ac.uk). This measures people paying, not the share of pages behind a paywall.

Older consent research is background only. The 2024 "Consent in Crisis" audit examined web sources underlying AI training corpora (arxiv.org); its restriction percentages concern training data collection and must not be used as inference-time PE retrieval rates. A 2025 IMC (Internet Measurement Conference) study of 130 self-declared bots found compliance highest for a crawl-delay directive and lowest for disallow-all (arxiv.org).

The limits are explicit. No public source reviewed here measures how often PE-relevant sources are unreadable to a given agent. The Web Almanac and Cloudflare figures are broad aggregates, and a practitioner's one-day robots file audit (2026-07-16) is point-in-time and limited to what a robots file can show (davidkemp.ai).

Case Studies and Real-World Examples

A fictional ledger walk-through (Hypothetical Example)

This case is fictional: "Northgate Fixtures Ltd" does not exist.

An agent is asked: does Northgate Fixtures Ltd have a corporate parent, and does it provide field-service maintenance? The mandate version is "v3, service-fit required, ownership must be identified".

1. **Ownership, registry source:** the agent discovers a registry page. The permission check records an approved API route. Retrieval returns 200, but the body is an HTML interstitial. The extraction stage flags the marker text, the failure category is "interstitial", and the disposition is "approved route": query the official API with a key. The answer "ownership not found" is not allowed at this stage.
2. **Ownership, API route:** the API returns a record with a controlling-person field. The excerpt is stored and the disposition is "accepted evidence", with the jurisdiction noted.
3. **Service fit, company site:** the page returns 200 with an empty application shell. The failure category is "dynamic content not rendered". The disposition is "analyst review", since the permitted interface cannot render it.
4. **Service fit, annual report:** a scanned PDF is retrieved but extraction yields no text ("extraction failure"). After a permitted OCR pass, the analyst reads a passage stating the company runs no maintenance division. The disposition is "read, contradicts", and the mandate question becomes a disqualifier hold.

In this walk-through, a naive pipeline would have closed question 1 as "not found" and question 3 as "no service-fit evidence", producing two false closures from two sources the agent had technically "visited". The ledger forces both into visible retrieval states. A practitioner's description of the general hazard is apt: a blocked source goes missing from an answer the way a case outside a database goes missing, and only knowing where the tool cannot look catches what it never saw (davidkemp.ai).

Implications and Future Directions

For research-operations leaders: the acceptance gate changes what a "no" is worth. A batch with high apparent accuracy can still hide false closures concentrated in the highest-risk questions. Reporting coverage by source class and question risk, and holding unresolved mandate disqualifiers, prevents an aggregate from masking them. A practitioner audit argues that exposure is subject-specific and that current private-markets intelligence is close to a dead zone for AI retrieval (davidkemp.ai), noting that private deal documents are private and public filings expose only part of the picture (davidkemp.ai). Both are attributed, dated opinions, not measured rates.

For vendors: trace visibility is the main open question. Agents that expose per-source status, final URL, content type and extraction outcome allow a full ledger; those exposing only a final answer allow a partial test. Procurement criteria can ask for these fields explicitly. Typed fetch errors returned inside a success response (platform.claude.com) and an option to restrict web search to cached or indexed results rather than live fetching (developers.openai.com) show the state space is already encoded in some vendor interfaces.

For the research community: the abstention and silent-failure literature mostly stubs tools or measures abstention on closed sets. An open fixture schema for web retrieval states would let teams compare traces. BrowseComp's authors report higher calibration error for browsing-enabled models (arxiv.org), a reminder that tool access can raise confidence without raising correctness.

Demand and uncertainty: The main uncertainties are PE-specific buyer demand and how much trace data vendors will supply. Policy also moves, so every ledger entry should carry a date and a rule reference, and replications should be rerun on current tool versions.

Frequently Asked Questions (FAQs)

How should teams evaluate AI research agents for private equity? Define the claim and source classes per question, run the eight-fixture pack against the agent, compute access coverage and false closure rate with Wilson intervals, and adjudicate a paired sample against an analyst or official API route. Judge the batch by per-question dispositions, not an overall score.

What belongs in an AI agent acceptance test checklist? Mandate version and required source classes per question; a five-stage source-attempt record; fixtures F1 to F8 with manifests; validation that challenge pages and snippets cannot pass as sources; paired review with reliability reporting; time-boxed retries within documented limits; coverage by source class and risk; and holds for unresolved disqualifiers. NIST's framework supports documenting test sets, metrics and tools in this way (nvlpubs.nist.gov).

How should a private equity deal sourcing AI agent be evaluated? On both discovery quality and access quality. Answer and citation benchmarks cover the first; this test covers the second. Ask for the agent's identity, its handling of robots rules and user-triggered fetches (developers.openai.com), and what trace it exposes.

Why do AI research agents fail on paywalled sources? Paywalls and login walls return authorization or interstitial states; HTTP 401 means the request lacks valid authentication credentials (www.rfc-editor.org). The right outcome is an authorization-required state routed to a licensed path, not a "not found" answer. Deep research tools document that connected accounts need access to the underlying data (help.openai.com).

Why does an agent say "not found" when the evidence exists, and is that a hallucination? Because retrieval and the answer are separate, and an HTTP success can carry an error in its body (platform.claude.com).

What vendor evaluation criteria matter for private equity AI tools? The procurement questions listed above, plus regulator expectations that firms evaluate generative AI tools before deployment (www.finra.org) and be able to conduct due diligence on third-party providers (www.iosco.org).

What is the difference between absence of evidence and not retrieved? Absence of evidence is a read state: the source was parsed and the field was missing. Not retrieved covers every earlier state: policy restriction, authorization, rate limit, challenge, unrendered content and extraction failure. Only the first can support an "unresolved" result.

How accurate does acquisition target research need to be? No public source reviewed here sets a threshold, so any threshold is an internal risk decision. The test supplies the measurement: false closure rate and coverage with intervals, by question risk.

What makes company research evidence-backed? Each accepted claim should link to a stored excerpt from a source actually obtained, with status, content type and timestamp. A cited URL alone does not show the agent read it.

Can robots.txt permission justify a workaround? No. RFC 9309 says its rules are not a form of access authorization and that the protocol is not a substitute for valid content security (www.rfc-editor.org). Equally, passing a robots check does not grant permission under a site's terms. Confirm permitted routes and terms separately.

Conclusion

The question this report addresses is whether an empty answer from a PE research agent is a finding or a gap. The answer depends on evidence the final text rarely carries: which sources were discovered, whether they were permitted, whether content was obtained, whether it rendered and extracted, and whether it was sufficient. The access-state ledger records those stages. The fixture pack tests the failure states under controlled conditions, including a genuinely absent fact so that legitimate unresolved results remain legitimate.

The decision rule follows. Accept a batch when each question has a disposition and a reason, when coverage and false closure rate are reported with their denominators and intervals, and when unresolved mandate disqualifiers are held rather than averaged away. Route policy-restricted questions to approved alternatives,

retry transient states within documented limits, and send ambiguous cases to analyst review. The test does not rank agents or endorse circumventing any control.

Limits remain. The strongest controlled evidence is a short-window preprint on simple pages, vendor documentation and site policies change, PE-specific base rates are unknown, and black-box vendors may expose too little to audit. Readers should consult their compliance and legal advisers on source terms and regulatory expectations in their jurisdictions.

Source links

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

1. https://arxiv.org/abs/2607.14447
2. https://platform.claude.com/docs/en/agents-and-tools/tool-use/web-fetch-tool
3. https://projecteuclid.org/journals/statistical-science/volume-16/issue-2/Interval-Estimation-for-a-Binomial-Proportion/10.1214/ss/1009213286.full
4. https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm
5. https://almanac.httparchive.org/en/2025/methodology
6. https://developers.cloudflare.com/cloudflare-challenges/challenge-types/challenge-pages
7. https://davidkemp.ai/blog/what-ai-legal-research-cant-read
8. https://arxiv.org/html/2607.14447v2
9. https://www.bain.com/insights/private-equitys-reality-check-gp-outlook-2026
10. https://www.deloitte.com/us/en/about/press-room/deloitte-survey-genai-in-mna.html
11. https://arxiv.org/abs/2504.12516
12. https://arxiv.org/abs/2506.21506
13. https://glossary.istqb.org/en_US/term/acceptance-testing
14. https://www.rfc-editor.org/rfc/rfc9309.html
15. https://www.fincen.gov/boi
16. https://www.fincen.gov/boi-faqs
17. https://www.gov.uk/guidance/people-with-significant-control-pscs
18. https://download.companieshouse.gov.uk/en_pscdata.html
19. https://developers.google.com/crawling/docs/robots-txt/robots-txt-spec
20. https://www.rfc-editor.org/rfc/rfc9110.html
21. https://www.rfc-editor.org/rfc/rfc6585.html
22. https://www.rfc-editor.org/rfc/rfc7725.html
23. https://developers.cloudflare.com/cloudflare-challenges/challenge-types/challenge-pages/detect-response
24. https://developers.cloudflare.com/rules/custom-errors/reference/error-page-types
25. https://vercel.com/blog/the-rise-of-the-ai-crawler
26. https://arxiv.org/abs/2607.19449
27. https://pmc.ncbi.nlm.nih.gov/articles/PMC3900052

28. https://developer-specs.company-information.service.gov.uk/guides/rateLimiting
29. https://www.sec.gov/os/accessing-edgar-data
30. https://www.sec.gov/search-filings/edgar-application-programming-interfaces
31. https://developer-specs.company-information.service.gov.uk/guides/authorisation
32. https://www.gleif.org/en/meta/lei-data-terms-of-use
33. https://api.gleif.org/docs
34. https://www.gleif.org/en/lei-data/access-and-use-lei-data/level-2-data-who-owns-whom
35. https://www.gleif.org/en/lei-data/access-and-use-lei-data/level-2-data-who-owns-whom/lei-roc-policy-on-level-2-data
36. https://e-justice.europa.eu/topics/registers-business-insolvency-land/beneficial-ownership-registers-interconnection-system-boris/general-information-finding-beneficial-owners_en
37. https://api.opencorporates.com/documentation/API-Reference
38. https://opencorporates.com/legal-information/self-service-api-terms-of-service
39. https://developers.openai.com/api/docs/bots
40. https://support.claude.com/en/articles/8896518-does-anthropic-crawl-data-from-the-web-and-how-can-site-owners-block-the-crawler
41. https://developers.google.com/crawling/docs/crawlers-fetchers/overview-google-crawlers
42. https://developers.google.com/crawling/docs/crawlers-fetchers/google-user-triggered-fetchers
43. https://docs.perplexity.ai/guides/bots
44. https://developers.cloudflare.com/bots/concepts/bot/verified-bots
45. https://developers.cloudflare.com/bots/reference/bot-verification/web-bot-auth
46. https://www.finra.org/rules-guidance/notices/24-09
47. https://www.finra.org/rules-guidance/guidance/reports/2026-finra-annual-regulatory-oversight-report/gen-ai
48. https://www.iosco.org/library/pubdocs/pdf/IOSCOPD823.pdf
49. https://www.fca.org.uk/firms/innovation/ai-approach
50. https://www.sec.gov/files/2026-exam-priorities.pdf
51. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf
52. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
53. https://www.ncbi.nlm.nih.gov/books/NBK557491
54. https://arxiv.org/abs/2506.09038
55. https://arxiv.org/html/2506.09038v1
56. https://arxiv.org/abs/1806.03822
57. https://pmc.ncbi.nlm.nih.gov/articles/PMC2550668
58. https://almanac.httparchive.org/en/2025/seo
59. https://blog.cloudflare.com/crawlers-click-ai-bots-training
60. https://blog.cloudflare.com/agentic-internet-bot-report
61. https://blog.cloudflare.com/radar-2025-year-in-review
62. https://blog.cloudflare.com/agentic-web

63. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2025/dnr-executive-summary
64. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2025/methodology
65. https://arxiv.org/html/2407.14933
66. https://arxiv.org/abs/2505.21733
67. https://developers.openai.com/api/docs/guides/tools-web-search
68. https://arxiv.org/html/2504.12516v1
69. https://help.openai.com/en/articles/10500283-deep-research

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.