Docket

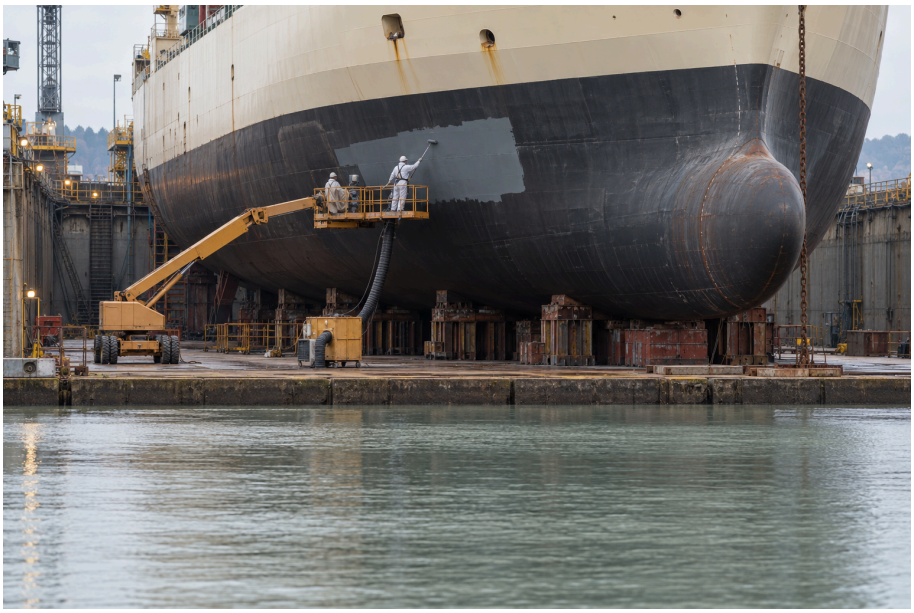
DOCKET / RESEARCH & PRACTICAL GUIDANCE

Auditing AI-Generated Company Research: A Review Checklist

> 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 · 2026-10-29 · ai generated research accuracy · audit ai research output · ai hallucination company data · verify ai generated citations

Original article: https://docket.capital/articles/auditing-ai-generated-company-research-nist



In this article

1. Why AI research needs an audit step
2. Failure modes in company research
3. What NIST AI 600-1 suggests
4. Retrieval is not proof
5. The reviewer's workflow
6. Sampling and escalation
7. Sign-off and record keeping
8. Limits of this checklist
9. Sources

Why AI research needs an audit step

A generated company profile reads the same whether it is right or wrong, which is why AI-assisted research needs a separate audit step rather than a quick read-through.

NIST's Generative AI Profile names the failure directly. Section 2.2 defines "confabulation" as a phenomenon in which generative AI systems "generate and confidently present erroneous or false content in response to prompts," and notes the behavior is colloquially called hallucination (NIST AI 600-1, July 2024). The same section warns that outputs "may also include confabulated logic or citations that purport to justify or explain the system's answer, which may further mislead humans into inappropriately trusting the system's output."

Two details in that section matter for deal teams:

- NIST says the dynamic is "particularly relevant when it comes to open-ended prompts for long-form responses and in domains which require highly contextual and/or domain expertise." A target profile is that kind of output.
- The profile describes automation bias as "excessive deference to automated systems" and says it can exacerbate confabulation risk (NIST AI 600-1). A citation next to a sentence raises a reader's trust whether or not anyone has opened the source.

The audit step breaks that link: a claim is accepted only when a reviewer has confirmed the cited material says it, about the right company, as of a known date.

Failure modes in company research

The research literature does not use one definition of hallucination, and a reviewer should not either. Ji et al.'s survey calls the most inclusive definition "the generated content that is nonsensical or unfaithful to the provided source content," and splits it in two: intrinsic hallucinations, where output "contradicts the source content," and extrinsic hallucinations, where output "cannot be verified from the source content" (Ji et al., Survey of Hallucination in Natural Language Generation).

For acquisition research, the useful categories are narrower and map to different review steps:

Failure mode	What it looks like in a target file	What catches it
Confabulated fact	A founding year, headcount or certification with no basis in any retained page	Open the citation; search the retained text for the value
Wrong-entity match	A true statement about a similarly named company in another state	Match legal name, address, domain and registry identifier
Stale fact	A 2019 "about us" claim presented as current	Record the page date and the date read; apply a recency rule per field
Non-existent or broken citation	A URL that returns an error, or a page that never contained the quoted text	Fetch or open the retained copy; confirm the quote verbatim
Misgrounded citation	A real, relevant-looking page that does not support the sentence attached to it	Read the cited passage against the claim, not just the page title
Unsupported summary	"Owner likely nearing retirement" with no stated basis	Treat inference as inference; require the underlying quotes

The misgrounded case deserves the most attention because it survives a quick check. Magesh et al., evaluating commercial AI legal research tools, define a response as "misgrounded if key factual propositions are cited but misinterpret the source or reference an inapplicable source," and count a response as hallucinated "if it is either incorrect or misgrounded" (Magesh et al., Hallucination-Free?). They argue these errors are "potentially more dangerous than fabricating a case outright, because they are subtler and more difficult to spot."

That study reports that the commercial legal research tools it tested "each hallucinate between 17% and 33% of the time" on its preregistered set of legal research questions (Magesh et al., arXiv 2024; the same figure appears in the published version, Journal of Empirical Legal Studies, 2025). Read that figure for what it is: a measurement of specific legal research products on legal queries. It is not a rate for company research, for any other vendor, or for any system you are reviewing. Its value here is the typology: a real citation can still be the wrong evidence.

Not every bad record is a hallucination: a faithful quote about the wrong company is an identity error, and the cause decides what to fix.

What NIST AI 600-1 suggests

NIST released the Generative AI Profile (NIST-AI-600-1) on July 26, 2024, as a companion to the AI Risk Management Framework, which NIST describes as "intended for voluntary use" (NIST, AI Risk Management Framework). The profile is a cross-sectoral profile that can be used "to govern, map, measure, and manage risks" (NIST AI 600-1), and its suggested actions carry IDs that follow those four functions (GV, MP, MS, MG).

The actions below are quoted verbatim from the PDF tables as read on 2026-10-28. NIST wrote them for generative AI generally; the right column is our mapping, not NIST's.

Action ID	NIST text (verbatim)	What it means for a research reviewer
MP-2.3-003	"Deploy and document fact-checking techniques to verify the accuracy and veracity of information generated by GAI systems, especially when the information comes from multiple (or unknown) sources."	Write the review procedure down; apply it hardest to claims drawn from directories, aggregators and unknown pages
MS-2.3-002	"Evaluate claims of model capabilities using empirically validated methods."	Test a vendor's accuracy claims on your own labeled sample before relying on them
MS-2.5-001	"Avoid extrapolating GAI system performance or capabilities from narrow, non-systematic, and anecdotal assessments."	A few good profiles on a demo call say little about a full target list
MS-2.5-003	"Review and verify sources and citations in GAI system outputs during pre-deployment risk measurement and ongoing monitoring activities."	Citation checks happen before adoption and on every batch after, not once
MS-2.5-005	"Verify GAI system training data and TEVV data provenance, and that fine-tuning or retrieval-augmented generation data is grounded."	Know which pages the system actually retrieved, and keep them
MG-4.3-002	"Establish and maintain policies and procedures to record and track GAI system reported errors, near-misses, and negative impacts."	Log every struck claim and its cause; the log is your evidence of how the system behaves

MS-2.5-003 is the core of this checklist. The phrase "ongoing monitoring" matters: a review done during a vendor trial does not cover a batch delivered months later, when sources, prompts or models may have changed.

Retrieval is not proof

Retrieval-augmented systems condition their answers on fetched documents. That moves the reviewer's problem rather than removing it.

The original retrieval-augmented generation paper framed the motivation plainly: for large pretrained models, "providing provenance for their decisions and updating their world knowledge remain open research problems," and the authors combined "pre-trained parametric and non-parametric memory" (in their case, a dense vector index of Wikipedia) to address it (Lewis et al., NeurIPS 2020). Retrieval gives the output a place it can point to. It does not guarantee the pointer is right.

Four things can still go wrong after a page is retrieved:

1. **The page is about a different entity.** Retrieval matches text, and two companies can share a name, a town or a product line.
2. **The page does not say what the claim says.** This is Magesh et al.'s misgrounded case: the source is real, the attribution is false (Magesh et al.).
3. **The page is out of date.** A cached or archived page can be faithfully quoted and still describe a company that has since been sold.
4. **The claim adds to the page.** Ji et al.'s extrinsic category: output that the source neither supports nor contradicts (Ji et al.).

Magesh et al. describe what checking actually requires: users must "click through to cited references, read and understand the relevant sources, assess their authority, and compare them to the propositions the model seeks to support" (Magesh et al.). Written about legal research, it describes the company research reviewer's job too.

The reviewer's workflow

Run the review at claim level, not record level. A record with mostly correct fields and one fabricated ownership claim is not "mostly right" for a screening decision; the ownership claim may be the only field that mattered.

Step 1. Split the output into atomic claims. One claim, one value, one citation. "Family-owned manufacturer in Ohio since 1994 with about 60 staff" (a fictional example) is four claims.

Step 2. Classify each claim. The class sets the evidence bar and the sampling tier.

Class	Examples	Minimum evidence
Identifier	Legal name, domain, address, registry number	Registry or primary record matching at least two identifiers
Status	Operating, independent, founder still in role	Dated primary source; recency rule applies
Ownership	Owner, family-held, sponsor-backed, recent sale	Primary source or filing; never inferred from silence
Financial	Revenue, headcount, locations	Source and its basis (stated, estimated, modeled), with date
Narrative	Positioning, customer mix, succession signals	Verbatim quotes that support each element of the summary

Step 3. Open the cited source. Prefer the retained copy the system used over today's live page; if they differ, record both.

Step 4. Confirm the quote and the date. Find the quoted text verbatim. Record the page's own date, if any, and the date read.

Step 5. Check entity identity. Match the source's company to the target on legal name, address, domain or registry number. A match on trade name alone is not a match.

Step 6. Test support, not topic. Ask whether the quoted sentence, read alone, establishes the claim. A markets page listing "marine and industrial" does not establish "aerospace."

Step 7. Assign a disposition and a cause. Supported, narrowed, unsupported, contradicted, wrong entity, or open gap. Record the cause (confabulation, identity, stale, misgrounded, extrinsic addition) so errors can be tracked under MG-4.3-002-style logging.

Illustrative example (fictional). A generated profile of "Tamsin Ridge Coatings," a fictional company, reviewed claim by claim. Company, sources and values are invented.

#	Claim as generated	Class	Cited source	Reviewer finding	Disposition
1	Legal name Tamsin Ridge Coatings, LLC	Identifier	State registry entry	Name, registered address and domain match	Supported

#	Claim as generated	Class	Cited source	Reviewer finding	Disposition
2	Family-owned, second generation	Ownership	Company "About" page	Page says "family-run since 1994"; nothing on generations	Narrowed to quoted text
3	Revenue about $18M	Financial	Business directory listing	Listing marks figure as an estimate, no date or basis	Unsupported as fact; keep only as labeled third-party estimate
4	Acquired by a PE sponsor in 2021	Ownership	Press release	Release concerns a same-name company in another state	Wrong entity; struck; escalated
5	Serves aerospace and marine customers	Narrative	"Markets" page	Page lists marine and industrial; no aerospace	Narrowed; aerospace struck
6	Owner approaching retirement	Narrative	None	No source	Open gap
7	Currently operating	Status	Homepage, read on review date	Live, recent posts dated this year	Supported, dated

Claim 4 alone would have changed the screening outcome, and it carried a working citation. Identity is checked before support for that reason.

Sampling and escalation

Whether to review every claim or a sample depends on what each claim decides. The arithmetic of sample sizes, acceptance numbers and confidence bounds is covered in Docket's audit sampling worksheet; this section covers how to tier the review and when to stop trusting a sample.

A practical tiering:

- **Tier 1, review 100%:** identifier, ownership and status claims that decide whether a company advances or is contacted. A wrong-entity error here can send outreach to the wrong owner.
- **Tier 2, review all used in a decision:** financial claims that drive a size screen. Financials shown for context only can be sampled.
- **Tier 3, sample:** narrative summaries, using a sample size and acceptance number written down before the review starts.

Stop rules should also be written in advance. Examples that do not require any assumed error rate:

- Any citation that does not exist, or a quote not found in its source, stops the sample. Treat the batch as unreviewed and move it to full review for that claim class.
- Any wrong-entity finding in Tier 1 triggers an identity re-check across the batch, because an identity error may come from a shared upstream step.
- Defects above the pre-set acceptance number expand the sample or reject the batch, per the worksheet.

- A change in model, prompt, source set or vendor resets the baseline. Earlier review results do not carry over (MS-2.5-001 and MS-2.5-003 above).

Do not turn review results into a system-wide accuracy percentage unless the sample was designed for it. A clean narrative sample says little about ownership claims, and a handful of spot checks is the "narrow, non-systematic, and anecdotal" assessment NIST warns against (NIST AI 600-1).

Sign-off and record keeping

The sign-off should let someone who was not in the room see what was checked, against what, and what remains open. The field-level structure for recording observations, conflicts and dispositions is laid out in Docket's claim ledger; the sign-off sits on top of it.

Sign-off template

Field	Entry
Batch identifier and size	Records and claims reviewed, by tier
System, version and run date	Tool or vendor, model or release if known, date the research ran
Reviewer and date	Name of the person accountable for the review; review date
Second reviewer	Required for any Tier 1 claim that changes a screening outcome
Sample design	Tier rules, sample sizes, acceptance numbers, stop rules triggered
Sources retained	Location of retained copies, with access dates
Dispositions	Count by disposition and by cause
Struck claims	List, with reason; kept, not deleted
Residual gaps	Open questions, owner of each, and whether they block outreach
Escalations	Items sent to deal lead or compliance, with outcome

Three record-keeping rules make the file useful later:

1. **Keep the struck claims.** A deleted error teaches nothing; a logged one shows which claim classes fail and why, the kind of record MG-4.3-002 describes.
2. **Keep the source copies.** Live pages change. A quote that cannot be reproduced from a retained copy is, for audit purposes, unsupported.
3. **Keep gaps visible.** An open "unknown" is a finding. Filling it with an inference to make the record look complete reintroduces the failure the review was meant to remove.

Limits of this checklist

This checklist is a working method, not an assurance standard, and passing it does not certify a file. It describes voluntary guidance and is not legal or compliance advice.

- **NIST AI 600-1 is voluntary.** The AI RMF is "intended for voluntary use," and NIST's framework page states: "The AI RMF 1.0 is being revised as part of the White House AI Action Plan" (NIST). As of October 28, 2026, the AI 600-1 PDF on NIST's site is still the July 2024 profile and the six action IDs above match it; IDs and wording may change in a future revision, so re-check them before citing them in a policy.
- **The literature is from adjacent domains.** Ji et al. survey natural language generation broadly; Magesh et al. measured legal research tools on legal questions; Lewis et al. evaluated retrieval on general knowledge-intensive NLP tasks. None measured company research, and none of their figures transfers to a system you are reviewing.
- **No accuracy claim is made for any system,** including Docket's. The only defensible accuracy statement about a research batch comes from a sample you designed and reviewed yourself.
- **Out of scope:** adversarial testing of research agents (for example, target websites that carry instructions aimed at the agent) is covered in Docket's prompt injection test suite, and sample-size calculation in the sampling worksheet.

Docket runs this kind of evidence check on target lists before outreach: its Audit agent re-reads retained pages with fetch tools denied, strikes anything it cannot quote and leaves gaps open; the method is described in Docket's evidence standards.

Sources

1. National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1)" (July 2024). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf. Accessed 2026-10-28.
2. National Institute of Standards and Technology, "AI Risk Management Framework" (web page, undated). https://www.nist.gov/itl/ai-risk-management-framework. Accessed 2026-10-28.
3. Ji, Z., Lee, N., Frieske, R., et al., "Survey of Hallucination in Natural Language Generation", ACM Computing Surveys 55(12), 1-38 (2023), https://doi.org/10.1145/3571730; quotations from the open arXiv version 2202.03629 v7 (July 14, 2024), https://arxiv.org/abs/2202.03629. Accessed 2026-10-28.
4. Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., Ho, D. E., "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools", Journal of Empirical Legal Studies 22(2), 216-242 (2025), https://doi.org/10.1111/jels.12413; quotations from the open arXiv version 2405.20362 (May 30, 2024), https://arxiv.org/abs/2405.20362. Accessed 2026-10-28.
5. Lewis, P., Perez, E., Piktus, A., et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" (NeurIPS 2020; arXiv 2005.11401). https://arxiv.org/abs/2005.11401. Accessed 2026-10-28.

Source links

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

1. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
2. https://arxiv.org/abs/2202.03629
3. https://arxiv.org/abs/2405.20362

4. https://doi.org/10.1111/jels.12413
5. https://www.nist.gov/itl/ai-risk-management-framework
6. https://arxiv.org/abs/2005.11401
7. https://doi.org/10.1145/3571730

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.