The Audit Gap

Organizational Vulnerabilities of Generative AI Adoption

Adam Roegiest
adam@roegiest.com
Zuva
Toronto, Canada

1 Introduction

Vertical software is a segment of the software market dedicated to meeting the needs of a specific industry or professional task. A prototypical example is the legal industry which has had dedicated software built for it, including time-tracking software, document management systems, and document review platforms (e.g., for electronic discovery, for contract review). This is in contrast to horizontal software that broadly applies across industries (e.g., word processors, spreadsheets, project management). More recently, we might consider the development and rise of conversational generative and agentic systems as a new form of horizontal software. These systems present a largely generic interface but can be tailored to the needs of individual users.

With this ability to "train" these generative systems (i.e., using `skills.md` prompts and tool use) and have them adapt to industry-specific use cases [2, 5, 10, 15, 16, 20, 21], this has led some to proclaim the death of vertical software [6, 11, 17, 20] and others offering balanced positions [18]. While these tools can be extremely effective for optimizing individual workflows,[1] our view is that these systems have not developed enough to make them effective replacements for vertical software. GenAI tools are designed for optimizing an individual's workflow rather than organizational ones, where there is a desire to ensure consistency of process across users. The ability to hyper-optimize for an individual is ideal for a single user but less so when an organization wants to provide guarantees around the work product (e.g., a diagnosis, a legal brief). Moreover, insight into what happened in these systems is not always clear and may be (unintentionally) obfuscated.

Vertical software, in contrast, often allows organizations to customize the software to meet their own desired processes and controls, whether through bespoke add-ons or a configurable interface. The more consequential difference, however, emerges in regulated and high-stakes domains, where vertical software is built around the controls and record-keeping that the industry itself requires. Because such software is designed for organizational control rather than individual productivity, auditability tends to follow as a downstream property (i.e., the system records what it and its users have done in a form the organization can later inspect). In the legal space, there are requirements around certification of document production (i.e., an eDiscovery lawyer certifies that the results are complete for the needs of the case) [19] and the American Bar Association's Opinion 512 places the onus on lawyers to understand the limitations of GenAI tools and to verify their output [1]. This is not a universal feature of vertical software, but it is characteristic of the domains where these tools displace skilled professional work, and it is increasingly a regulatory expectation (e.g., Article 12 of the EU AI Act [7] and echoed in WHO guidance for health applications [22]). This audit trail is what lets an organization verify that the desired actions took place and gives it confidence that outputs meet its guidelines for quality and regulatory compliance.

What an organization needs for auditing is not a faithful window into the model's internal computation but a complete, reliable, and searchable record of the consequential steps in a task. For example, the queries issued, the sources retrieved, and what was included or excluded and why. Reasoning traces have largely been framed in terms of faithfulness and whether the visible "thinking" corresponds to the computation that actually produced the output [12–14]. For auditing purposes, an organization does not need to know what the model "thought;" it needs a meaningful record of what the system did that it can inspect and verify against its own policies. Our argument is not that these traces misrepresent any supposed model cognition, but that they are the only audit artifact currently available and that they fail as one.

As we briefly demonstrate in Section 2, systems often surface their reasoning to the user through interactive elements in the interface. As a record, this artifact is deficient on every dimension above. It is incomplete: what is shown is a summary presented to the user rather than the full process the model carried out [12], and it omits the specifics of decisions in favor of high-level description. It is not attributable: the trace rarely records which sources were consulted or how they shaped subsequent steps, leaving the user to infer this from the final citations. And it is not searchable, as we discuss below. For low-impact activities this is tolerable (e.g., not knowing why a particular result was used is rarely critical to an academic literature review), but for high-impact activities an incomplete, unattributable record can materially affect the outcome and pose risks to the end user. For example, a relevant case might be omitted from a generated legal analysis despite being recorded as present in a set of search results. If the classification logic is not present, and in our experience it is not, then it would be easy to overlook this omission when reviewing the intermediate results. Such aspects may contribute to the non-trivial number of hallucinations in court documents[4], ranging from fake citations to false quotes to misrepresentation of facts. That is, if any reasoning is present, it may not be of particular use to the user in determining the quality of the system's outputs and so the user must trust the system to be correct.

One might argue that the system can simply be interrogated after the fact once an issue is noticed. But this substitutes a freshly generated explanation for a contemporaneous record, and the two are not equivalent. Since these systems tend toward sycophancy, an interrogated system is likely to acknowledge the supposed mistake, supply a plausible reason for it, and express a willingness to correct

[1] Incautious use has also been problematic in practice [3, 8, 9].

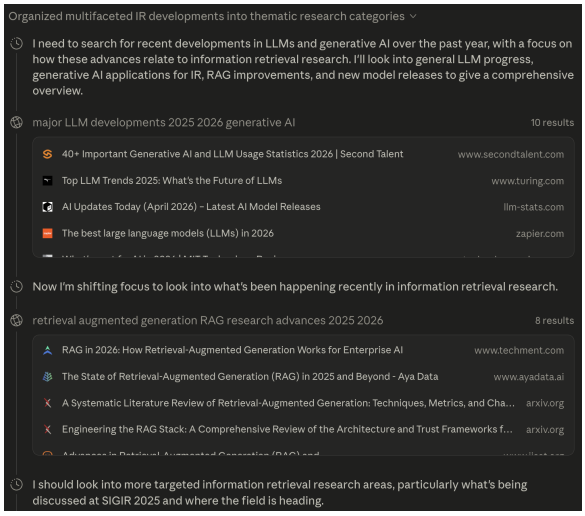

Figure 1: A snippet of Claude Opus 4.7's "thinking" as presented to a user. The full conversation can be found at https://claude.ai/share/6797200c-a672-487f-bf39-6872e768a6f4.

the error. However, that reason is produced in the moment and is not anchored a record of what was actually done, so it need not correspond to the decision the system made earlier in the process. We do not claim the intermediate "thinking" (or reasoning) faithfully reflects the model's internal state; we claim only that a decision was made, that it went unrecorded, and that a sycophantically generated explanation is an unreliable reconstruction of it. The organization is left with a trace that is an incomplete record and a subsequent explanation that is unreliable. This outcome is not one that readily facilitates an audit or even subsequent correction to processes (e.g., any instruction tuning).

Finally, the "thinking" of these systems is not searchable. Even when a trace does record a consequential step, users and organizations cannot systematically query across interactions to verify that a defined process was followed. Instead, a user must exhaustively review intermediate output by hand to confirm, for example, that the procedure specified in a skills.md was applied. This is the searchability failure named above, and it is what makes the deficiency an organizational problem rather than an individual inconvenience, as verification does not scale.[2]

2 Examples

Figures 1 and 2 depict snippets of the generated "reasoning" presented to the user for the prompt: "Summarize all the developments that have happened in the last twelve months with respect to LLMs and generative AI, especially as it relates to Information Retrieval research." We used a literature-review-flavored prompt as an illustrative example to avoid the nuances of a specialized task. OpenAI took substantially longer to formulate its response due to the increased thinking time allowed by the organizational plan, but the

[2]OpenAI has itself discussed issues around monitoring reasoning models and the challenges therein [13, 14].

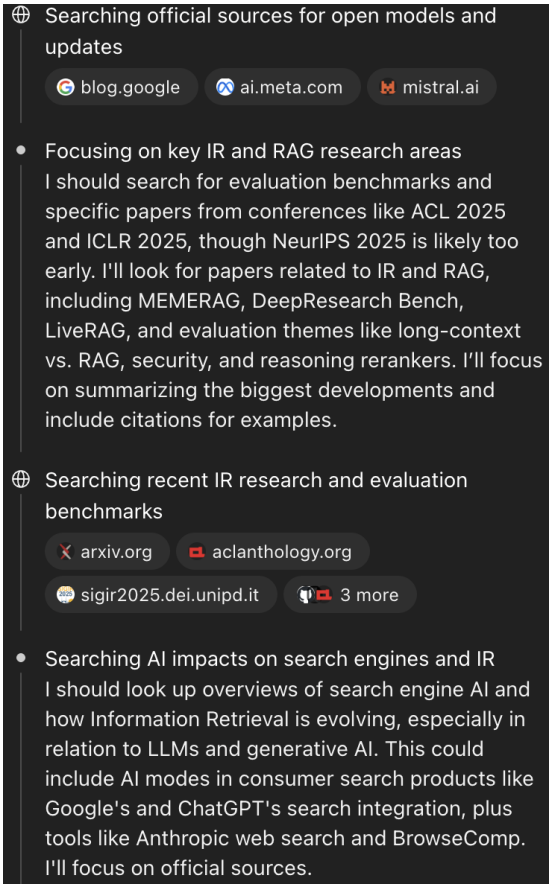

Figure 2: A snippet of OpenAI ChatGPT Pro's "thinking" (later in the process) as presented to a user demonstrating more verbose reasoning than Claude. Full conversation is unable to be shared due to workspace settings.

structure between the two is largely the same: each states its understanding of the task, issues a series of searches to map recent AI developments as they relate to IR, and then composes the summary. We read both traces against the three properties an audit record requires (completeness, attributability, and searchability) and find each one lacking on this deliberately low-stakes task.

Completeness. For the writing stage, Claude surfaces a single summarizing thought, whereas ChatGPT produces over twenty thoughts concerning content, formatting, and date-specific information. This demonstrates that the granularity of the record varies not only across systems but across phases of the same task, with no guarantee that consequential steps are the ones captured. The record is also explicitly a summary rather than the full process the model carried out [12]. In particular, the user sees what the interface chose to show and not what was done. This highlights that the reasoning traces will not correspond to a consistent, complete set of decisions but will vary across task stages.

Attributability. Neither system describes which search engine was used, which returned results were judged relevant, or how those results shaped subsequent searches. Claude's trace at least

lets one infer the query from its phrasing, whereas ChatGPT reports only the intent of a search, leaving the connection to prior steps implicit. In particular, ChatGPT notes that it drew on 440 sources, while Claude leaves the count for the reader to reconstruct from the log, and neither records which of those sources actually contributed to the summary. ChatGPT's search tactics, moreover, shift between navigational look-ups of specific vendors and broad exploratory queries, yet how it settled on particular vendors and omitted others goes unrecorded. The user is left to infer attribution from the interaction history and the final citations, which is manual reconstruction that an audit record is supposed to make unnecessary.

Searchability. This example is a single interaction, so the inability to search across many "thinking" histories is not directly visible here. But even within one trace, nothing is structured for retrieval: to confirm what, if any, procedure was followed, a reviewer must read the intermediate output in full rather than query it. What is already tedious for one low-stakes interaction does not scale to the volume of high-impact work an organization would need to verify process adherence on a regular basis. Taken together, even on a generic task the surfaced reasoning fails as an audit record on all three counts. On a high-impact task, where an omitted clause or an unrecorded decision changes the outcome, the same failure becomes consequential, and it is already documented in practice.

A concrete illustration appears in the Vals Legal AI Report [20], in which tools were asked to extract any clause relating to a most-favored-nation (MFN) provision. The contract contained no clause with that heading, but it did contain the requisite language (i.e., treatment "no less favorable" than that afforded any other customer). That phrasing is an easily searchable signal of an MFN provision. The lawyer baseline and one tool retrieved the clause and the remaining tools either returned an irrelevant clause or reported finding nothing. The failure itself is unremarkable; what matters for our argument is that the systems offered no account of it, surfacing neither what they searched for nor why the relevant language went unidentified. While the study's practitioners hypothesized that these failures were due to a misalignment between the system's expectation (e.g., a clear "Most Favored Nation" heading) and reality, there was no way to verify that. This is the a simplified example of the audit gap, where there is a determinable, consequential omission with no contemporaneous record of the decision that produced it, so that even a careful external explanation remains a reconstruction the organization cannot easily confirm.

There are, unfortunately, no obvious examples of the the audit gap in practice. The horizontal tools being positioned as replacements for vertical software are themselves nascent: Claude for Legal, for example, was launched only in May 2026, building on a Claude Cowork legal plug-in from February 2026. The audit gap we describe is therefore not a transitional artifact to be assumed away, but a property of systems being developed and adopted faster than their organizational impact can be characterized.

3 Conclusion

We argue that for GenAI systems to be of value to professionals, they must produce an auditable record of their consequential intermediate decisions (i.e., what was retrieved, used, and excluded) rather than a faithful summary of their reasoning. That record must be reliable and searchable so that policies and guidelines can be verified at scale. Absent this, organizations risk poor decisions made due to overconfidence in systems that tend toward sycophancy.

References

[1] American Bar Association Standing Committee on Ethics and Professional Responsibility. 2024. Formal Opinion 512: Generative Artificial Intelligence Tools. American Bar Association. https://www.americanbar.org/content/dam/aba/administrative/professional_responsibility/ethics-opinions/aba-formal-opinion-512.pdf Accessed 2026-06-03.

[2] Kent Bennett, Byron Deeter, Mike Droesch, Maha Malik, Sam Bondy, Brian Feinstein, Sameer Dholakia, Caty Rea, Alex Yuditski, and Aia Sarycheva. 2024. Part I: The future of AI is vertical. https://www.bvp.com/atlas/part-i-the-future-of-ai-is-vertical Atlas essay; accessed 2026-04-20.

[3] Moritz Blum. 2023. ChatGPT Produces Fabricated References and Falsehoods When Used for Scientific Literature Search. *Journal of Cardiac Failure* 29, 9 (2023), 1332–1334. doi:10.1016/j.cardfail.2023.06.015 Epub 2023-07-03.

[4] Damien Charlotin. 2026. AI Hallucination Cases. Online database. https://www.damiencharlotin.com/hallucinations/ Database of legal decisions addressing hallucinated content from generative AI; updated daily; accessed 2026-06-03.

[5] Jonathan H. Choi, Amy B. Monahan, and Daniel Schwarcz. 2024. Lawyering in the Age of Artificial Intelligence. *Minnesota Law Review* 109, 1 (Nov. 2024), 147. doi:10.24926/265535.4225

[6] Citrini and Alap Shah. 2026. *The 2028 Global Intelligence Crisis: A Thought Exercise in Financial History, from the Future.* Citrini Research. https://www.citriniresearch.com/p/2028gic Accessed 2026-04-20.

[7] European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng Official Journal of the European Union; relevant discussion in Article 12 on record-keeping.

[8] Daniel E. Ho. 2024. Hallucinating Law: Legal Mistakes with Large Language Models are Pervasive. https://hai.stanford.edu/news/hallucinating-law-legal-mistakes-large-language-models-are-pervasive Stanford HAI news page; contributors: Matthew Dahl, Varun Magesh, and Mirac Suzgun; published 2024-01-11; accessed 2026-04-20.

[9] Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Chanwoo Park, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten Sap, Yulia Tsvetkov, Paul Liang, Xuhai Xu, Xin Liu, Chunjong Park, Hyeonhoon Lee, Hae Won Park, Daniel McDuff, Samir Tulebaev, and Cynthia Breazeal. 2025. Medical Hallucinations in Foundation Models and Their Impact on Healthcare. arXiv:2503.05777 [cs.CL] https://arxiv.org/abs/2503.05777

[10] Lauren Martin, Nick Whitehouse, Stephanie Yiu, Lizzie Catterson, and Rivindu Perera. 2024. Better Call GPT, Comparing Large Language Models Against Lawyers. *arXiv preprint arXiv:2401.16212* (2024). doi:10.48550/arXiv.2401.16212

[11] Ben Murray. 2026. *The SaaSpocalypse: AI Agents, Vibe Coding, and the Changing Economics of SaaS.* The SaaS CFO. https://www.thesaascfo.com/the-saaspocalypse-ai-agents-vibe-coding-and-the-changing-economics-of-saas/ Posted March 10, 2026; accessed 2026-04-20.

[12] OpenAI. [n. d.]. Reasoning Models. https://developers.openai.com/api/docs/guides/reasoning OpenAI API documentation.

[13] OpenAI. 2025. Detecting Misbehavior in Frontier Reasoning Models. https://openai.com/index/chain-of-thought-monitoring/ OpenAI publication.

[14] OpenAI. 2025. Evaluating Chain-of-Thought Monitorability. https://openai.com/index/evaluating-chain-of-thought-monitorability/ OpenAI research publication.

[15] OpenAI. 2026. Introducing OpenAI for Healthcare. https://openai.com/index/openai-for-healthcare/ Product page, published 2026-01-08; accessed 2026-04-20.

[16] Ashwin Ramaswamy, Alvira Tyagi, Hannah Hugo, Joy Jiang, Pushkala Jayaraman, Mateen Jangda, Alexis E. Te, Steven A. Kaplan, Joshua Lampert, Robert Freeman, Nicholas Gavin, Ashutosh K. Tewari, Ankit Sakhuja, Bilal Naved, Alexander W. Charney, Mahmud Omar, Michael A. Gorin, Eyal Klang, and Girish N. Nadkarni. 2026. ChatGPT Health performance in a structured test of triage recommendations. *Nature Medicine* (2026). doi:10.1038/s41591-026-04297-7 Published online 2026-02-23.

[17] Zack Shapiro. 2026. *The Claude-Native Law Firm.* https://x.com/zackbshapiro/status/2027389987444957625 Accessed 2026-04-20.

[18] Angela Strange and James da Costa. 2024. Vertical SaaS: Now with AI Inside. Andreessen Horowitz. https://a16z.com/vertical-saas-now-with-ai-inside/

[19] U.S. Courts. 2007. Federal Rules of Civil Procedure, Rule 26(g): Signing Disclosures and Discovery Requests, Responses, and Objections. Legal Information Institute, Cornell Law School. https://www.law.cornell.edu/rules/frcp/rule_26 As amended; accessed 2026-06-03.

[20] Vals AI. 2025. *Vals Legal AI Report.* Industry report. Vals AI. https://www.vals.ai/industry-reports/vlair-2-27-25 Last updated 2025-02-27.

[21] Vals AI. 2025. *VLAIR - Legal Research.* Industry report. Vals AI. https://www.vals.ai/industry-reports/vlair-10-14-25 Legal Research Report.

[22] World Health Organization. 2024. *Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models.* Technical Report. World Health Organization, Geneva, Switzerland. https://www.who.int/publications/i/item/9789240084759 WHO guidance; publication page dated 25 March 2025, document citation metadata indicates 2024.