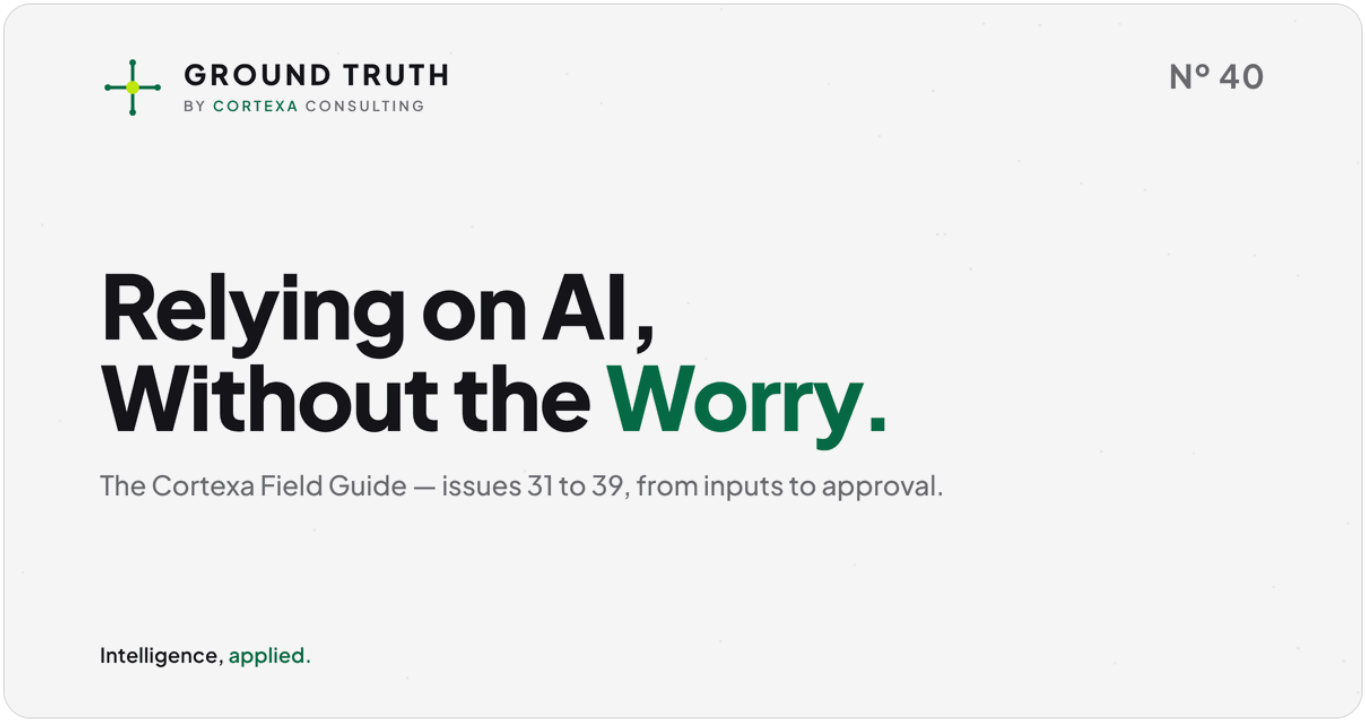
FIELD GUIDE

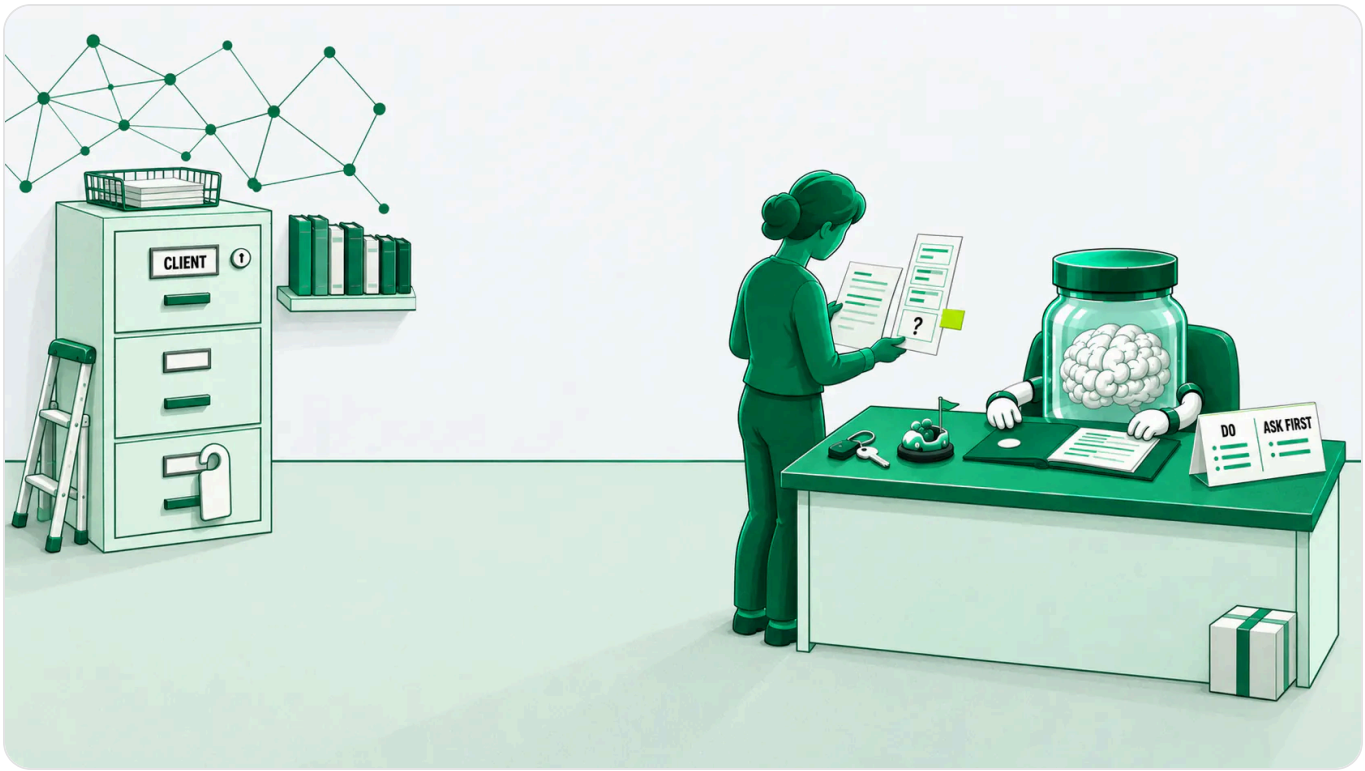
Relying on AI, Without the Worry.

Before a team relies on AI, run four checks. Look at the inputs and the outputs, who decides the next step, and who approves the work before it goes out.



*Intelligence, **applied.*** · November 19, 2026 · 20 min read · **www.cortexa-consulting.ai/insights/relying-on-ai-without-worry**

The first three Field Guides covered what Artificial Intelligence (AI) is (AI, Without the Hype[1]), how to build with it (Building With AI, Without the Guesswork[2]) and what happens once the work ships (Shipping AI, Without the Surprises[3]). This fourth one covers a step most teams have already taken without announcing it: relying on AI every day, for real client work. Nine explainers sit behind these chapters. Each started as a question someone asked out loud, and nearly all of them come down to four checks to run before a team relies on the results. First, the inputs: what you give the AI (documents, data and prompts) and which account it goes into. Second, the outputs: what the AI produces, and whether it is correct and yours to use. Third, the next step: who decides what happens next, whether that is a person, a fixed script or the AI itself. Last, approval: who checks the work and approves it before it reaches a client or the public. Each chapter names its check, shows what the evidence supports, and links to the explainer with the detail.

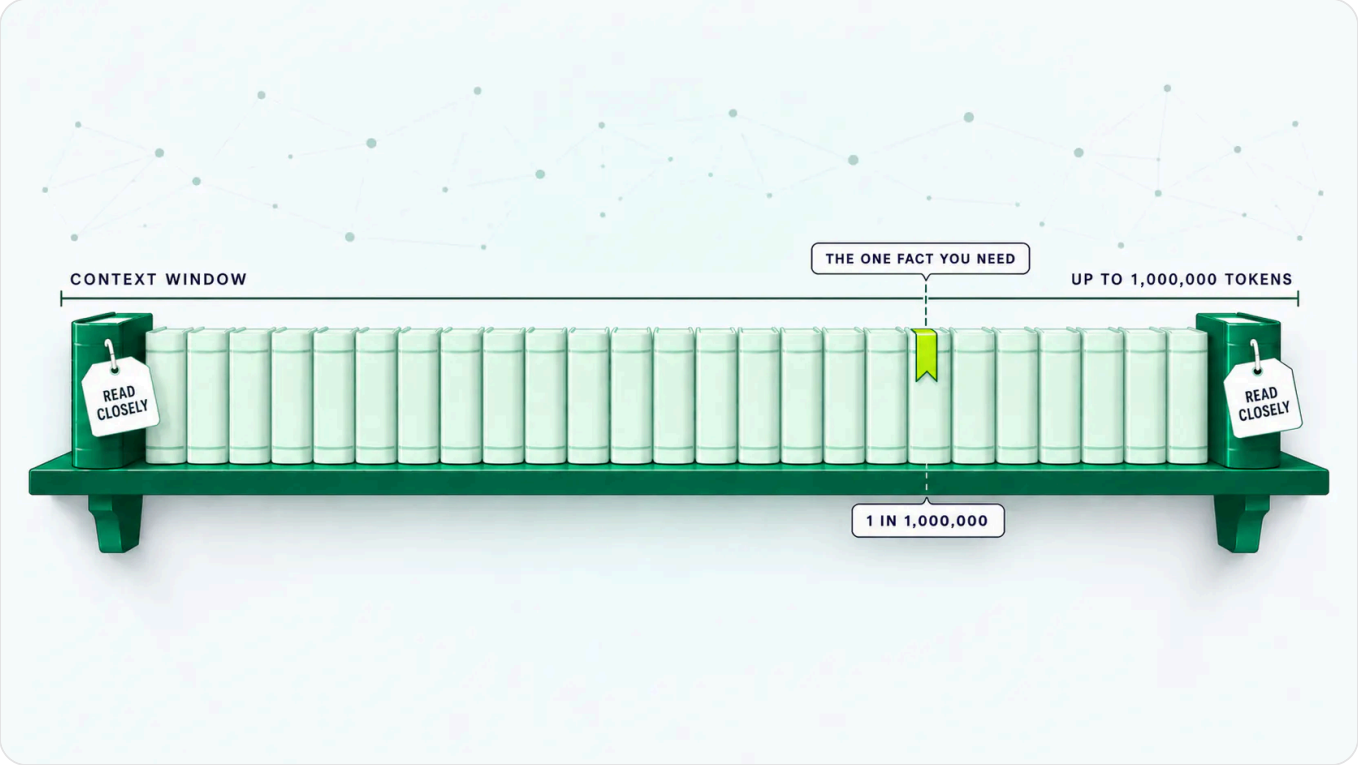


The four checks at one desk: what goes in, what comes out, who decides the next step, and a person who compares the work with its source before it leaves.

1. What it can see

Every check starts with what the model is looking at. Its context window[4] is all the text a language model can reference when generating a response, including the response itself[13]. That window has grown fast. Anthropic's Opus and Sonnet models now carry a 1M-token

context window[14], enough for a long brief, a full style guide and a stack of old correspondence in one call. The size invites a shortcut: paste in everything and let the model sort it out.



Everything fits on the shelf. The model reads the ends most closely, and the one fact you need may be sitting in the middle.

- A model can hold more than it reads closely. Anthropic's engineers write that as the number of tokens in the context window increases, the model's ability to accurately recall information from that context decreases[15].
- Position matters. Researchers found that performance is often highest when relevant information occurs at the beginning or end of the input context[16], and that it drops when the fact sits in the middle.
- The advertised number is a ceiling. NVIDIA's RULER benchmark tested seventeen models that all claimed at least 32,000 tokens and found only half of them can maintain satisfactory performance at the length of 32K[17].

CORTEXA'S READ

A window size tells you what a model can hold. Whether it reads the part that matters is something only the client's own documents can test. Put the fact that has to land near the start or the end, or fetch it fresh when it's needed, and measure before a workflow depends on it.

2. Where the data goes

The next question is where a prompt ends up once it leaves the screen, and the answer depends on the account more than the brand. On the free ChatGPT app, what you type can help train the model until someone switches that off in Data Controls[18]. On ChatGPT Business and Enterprise the default flips: OpenAI does not use business customer content from ChatGPT Enterprise and ChatGPT Business to train its models[19]. Kept out of training is a different fact from deleted. Enterprise retention has a minimum of 90 days[19], and the API keeps abuse-monitoring logs retained for up to 30 days[20] unless a Zero Data Retention agreement removes them. The client data explainer[5] walks through each tier.



By default, Business, Enterprise and the API keep your prompts out of training, but they still store them for a while. "Not used to train" is a different fact from "not kept."

But the harder question is which account people are using at all. Shadow AI[6] is staff using AI tools nobody approved, and it is common. Microsoft and LinkedIn reported that 78% of AI users are bringing their own AI tools to work[23]. It is a self-reported vendor survey, so read the number loosely. Still, the direction is hard to argue with.



The same client folder, two places to copy it. The shop across the street works on its own terms; the copier in the office is ready, and closer.

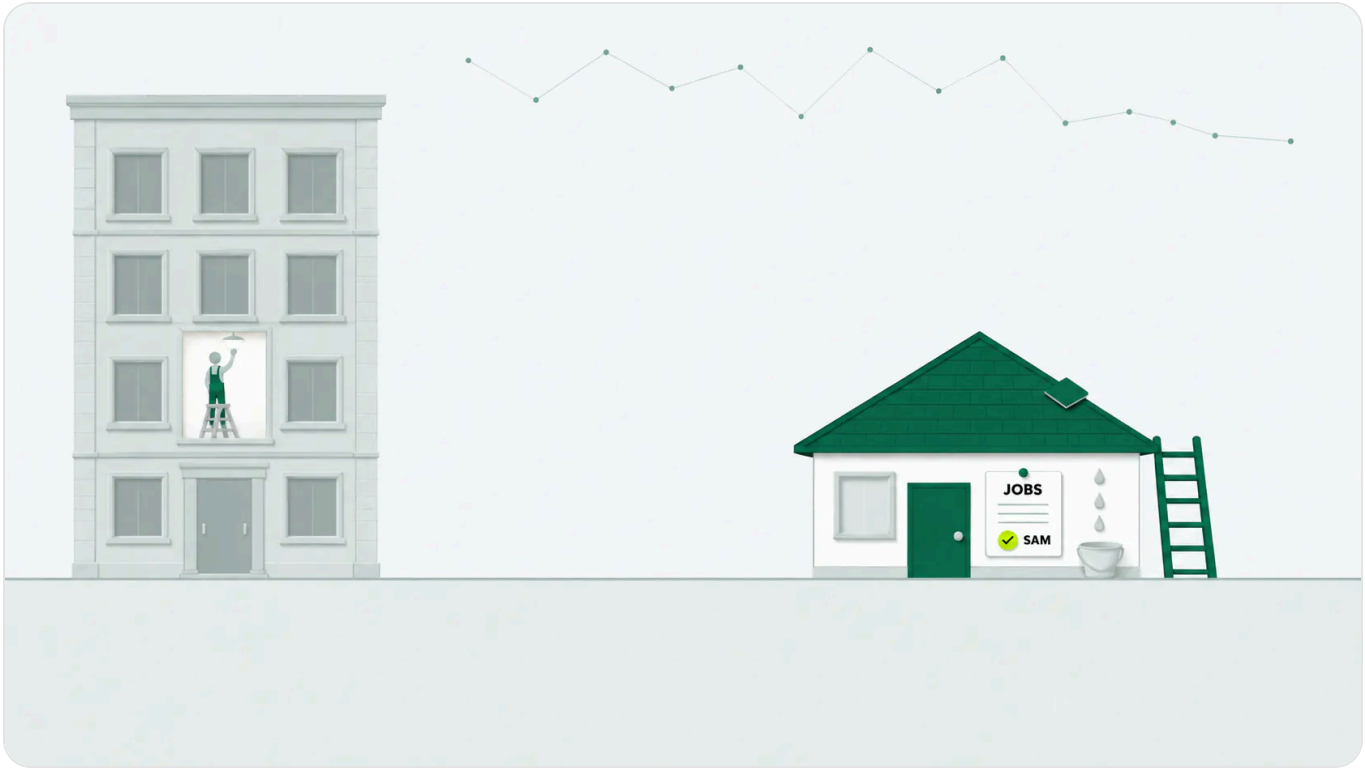
- The risk sits in the input. The Open Worldwide Application Security Project (OWASP) lists sensitive information that includes financial data, credentials and trade secrets[24], which is exactly what lands in a rushed prompt.
- Other vendors draw the same line. Anthropic keeps its commercial products out of training by default[21] and says it will automatically delete inputs and outputs within 30 days of receipt or generation[22].
- A ban tends to hide the habit. When most people who use AI already bring their own tools, banning them moves the use out of sight, where nobody can ask what data goes into those tools.

CORTEXA'S READ

A settings page is a default that a vendor can change with the next update. The signed Data Processing Agreement is what binds it. So the policy that protects a client fits on one page: which data never goes into an unapproved tool, which few tools are approved on business plans, and one named person who answers questions. Ask the team what they already use before you write it.

3. Whose machine runs it

Some clients ask whether to skip the vendors and run an open-weight model[8] themselves. Open weight means you get the learned numbers that make up the finished model and can run it on hardware you control, so prompts never leave your environment. The quality gap has shrunk on one public measure. In January 2024 the leading closed model outperformed the top open-weight model by 8.04%[25] on the Chatbot Arena leaderboard, and by February 2025 this gap had narrowed to 1.70%[25]. So the real change is who does the work a vendor's API quietly did for you: hosting, updates, and the layers that keep output safe and on brand.



In the apartment, the building fixes the light. In a house you own, the leak is yours, so the jobs list needs a name.

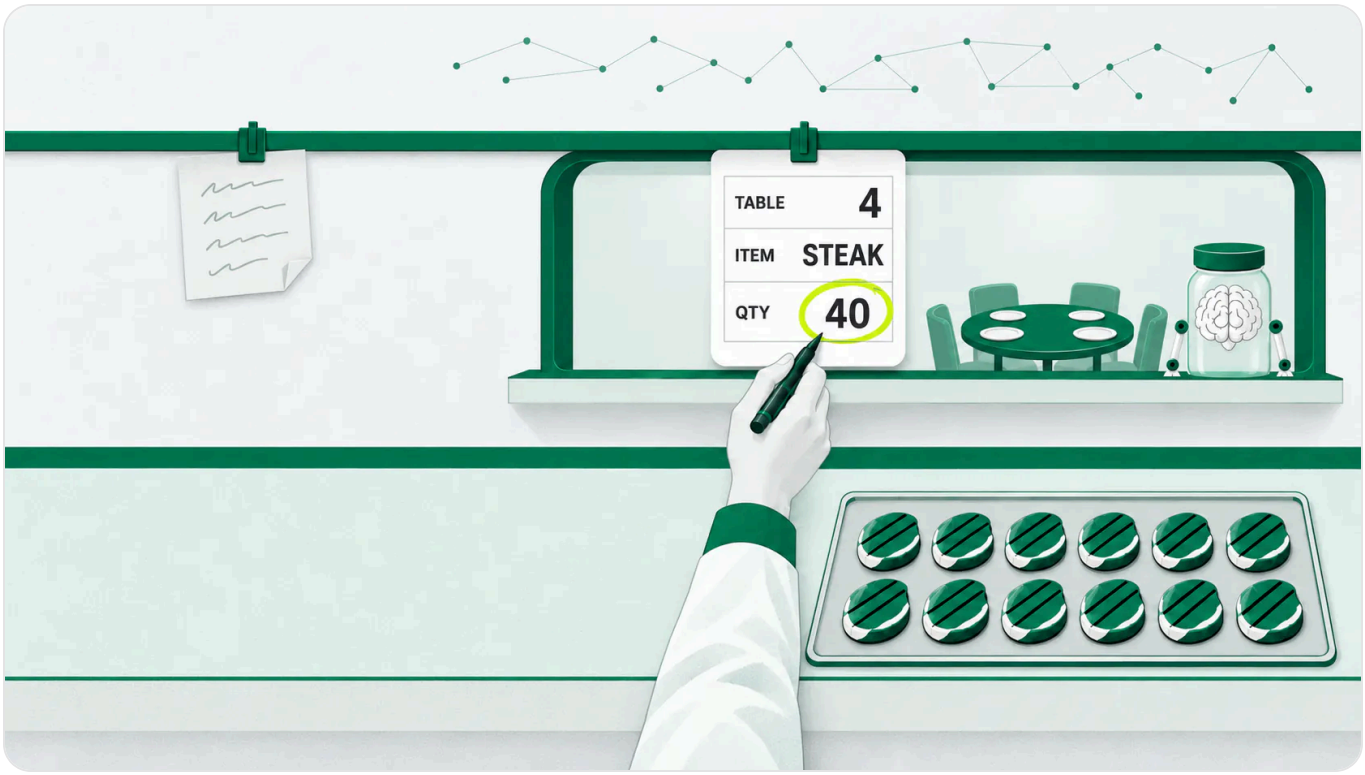
- Read the license before the benchmarks. OpenAI's gpt-oss models ship under the Apache 2.0 license, with a minimal usage policy[26], while Meta's Llama license asks products past 700 million monthly active users in the preceding calendar month[27] to request a license.
- The hardware bar is lower than many assume. The smaller gpt-oss model runs within 16GB of memory[26].
- Open does not mean exempt. The EU AI Act's open-source carve-out covers only some provider duties, and it shall not apply to general-purpose AI models with systemic risks[28].

CORTEXA'S READ

Treat it as a purchasing decision. A closed model is a service you rent; an open-weight one is equipment you own, and equipment needs someone to run it. Keeping data in-house and tuning for a narrow job are reasons that survive scrutiny. “It's free” rarely does, because the bill moves to people and servers.

4. The shape of the answer

Once work comes back, software often has to read it. Structured output[9] makes a model answer in a fixed format, usually JavaScript Object Notation (JSON), against a schema that names every field. Function calling applies the same idea to actions: the model fills in a request for a tool, and your application executes each call[30]. The format problem is largely solved. OpenAI says its feature means you don't need to worry about the model omitting a required key[29].



The ticket guarantees the boxes. Someone still has to check the number.

- The shape is guaranteed; the content is not. OpenAI's own guide says Structured Outputs can still contain mistakes[29]. A tidy record with the wrong budget in it is still wrong.
- A form with no honest empty answer invites invention. The model tries to fit the schema even when the input doesn't, which can result in hallucinations[29]. Give it a way to say "unknown."
- Whatever reads the output next should check it. OWASP names insufficient validation, sanitization, and handling[31] of model output before it reaches other systems as a risk.

CORTEXA'S READ

A clean, valid record looks trustworthy, so nobody reads it, and that is where teams get burned. Check the values a person would check: the budget is a positive number, the order exists. Then decide in advance what happens when the answer comes back "unknown" or the model declines.

5. Who picks the next step

Vendors now call almost everything an agent, and one question sorts the labels[12]: who decides the next step? With a chatbot or a copilot, a person does. With a workflow, a route someone

wrote in advance does. Anthropic defines workflows as systems where LLMs and tools are orchestrated through predefined code paths[32]. With an agent, the model decides. In Anthropic's words, LLMs dynamically direct their own processes and tool usage[32].



Same fairground, two wheels. One is for show. The one that steers comes with a wall and a stop button.

- Autonomy comes with brakes by design. Anthropic recommends stopping conditions (such as a maximum number of iterations)[32], and OpenAI's guide says that on failure an agent can halt execution and transfer control back to the user[33].
- Whether a person looks first is a setting someone chose. Microsoft says its agents can be configured to react to events or triggers without human input[34].
- Simpler is often enough. Anthropic advises finding the simplest solution possible, and only increasing complexity when needed[32].

CORTEXA'S READ

An agent earns the name when the model chooses the steps and can act in other systems. That is exactly when you need to know where it stops and who is watching. If a workflow does the client's job, recommend the workflow. Calling it an agent adds a price and no capability.

6. Who checks the work

“A person reviews everything before it goes out” is the promise most agencies make about AI-assisted work, and the review explainer[10] is about keeping it true. The failure has a name, overreliance. A Microsoft Research review of about 60 papers describes it as users accepting incorrect AI outputs[35]. Asking the AI to explain itself helps less than you would hope. In one preprint, explanations increased the chance that humans will accept the AI’s recommendation, regardless of its correctness[36].



The signature looks the same either way. Only one box was opened.

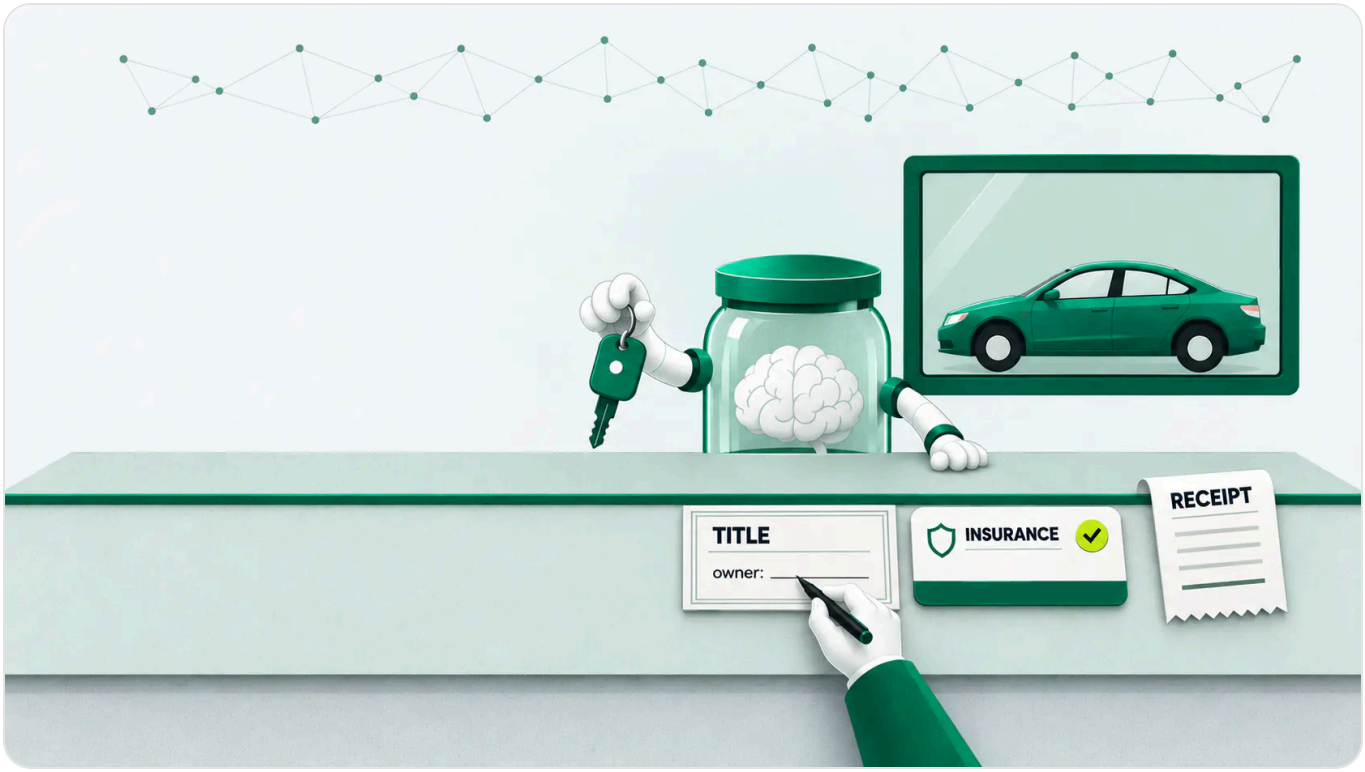
- Authority matters as much as attention. The EU AI Act's Article 14 says people overseeing a high-risk system must be able to decide not to use[37] it, and to stop it. A reviewer who cannot reject the work is only watching it.
- Put approval on the actions with consequences. OWASP advises human-in-the-loop control to require a human to approve high-impact actions[38].
- None of these sources says how much review is enough. That depends on the stakes and the error rate, so treat a vendor's fixed sampling number as theirs.

CORTEXA'S READ

Three questions sort a real review from a signature. What does the reviewer see besides the draft? How long do they have? What happens when they say no? If nobody can answer, the process has a signature where the review should be.

7. Who owns what it made

Generated images are where relying on AI turns into paperwork. An AI-made picture can usually run in a paid ad, and the images explainer[7] splits the question into who owns it, who pays if someone objects, and where it came from. On ownership, the U.S. Copyright Office concluded that prompts alone do not provide sufficient human control to make users of an AI system the authors of the output[39]. A federal appeals court held that the Copyright Act requires all eligible work to be authored in the first instance by a human being[40]. What people add still counts: protection reaches a human author's original expression even if the work also includes AI-generated material[39].



The image is the car. The paperwork is what makes it safe to drive.

- Indemnity exists, with strings. Microsoft ties its commitment to the customer using the content filters and other safety systems[41] built into the product, and Google Cloud's only

applies if you didn't try to intentionally create or use generated output to infringe the rights of others[42].

- Platforms label what they can detect. Meta applies an "AI info" label to ads created or edited using third-party AI tools[43] when it finds the industry signals, so a quiet AI image may not stay quiet.

> **CORTEXA'S READ**
>
> The risk in an AI image is rarely the image. It is the missing paperwork around it. Use AI to explore and to draft, let a person finish the asset that carries the client's name, and keep a receipt file with the tool, the version, the prompt and the date beside the final.

8. The person on the other end

Some AI work reaches a customer directly, as a message tailored to them. Personalization[11] helps until it feels like spying, and a peer-reviewed study shows where that line sits. When a firm collected data openly, more personalized ads drew higher click-through intentions[44]. When it gathered the same data covertly, the pattern flipped, and in the field study click-through rates dropped sharply once people realized their information had been collected without their consent[44]. It is a 2015 study of Facebook ads, so treat it as a pattern to test in your own audience.



Same hotel, same hallway. The pillow was asked for. The search was not.

- AI raises the stakes by making more data usable. A model can guess things a customer never typed in[11], and a guess about someone is information they never handed over.
- The ad platforms already fence off sensitive ground. Google says audiences an advertiser builds from its own data cannot be used if promoting in sensitive interest categories[45].
- In the European Union (EU), people can say stop. The General Data Protection Regulation (GDPR) gives them the right to object at any time[46] to direct marketing, profiling included.

CORTEXA'S READ

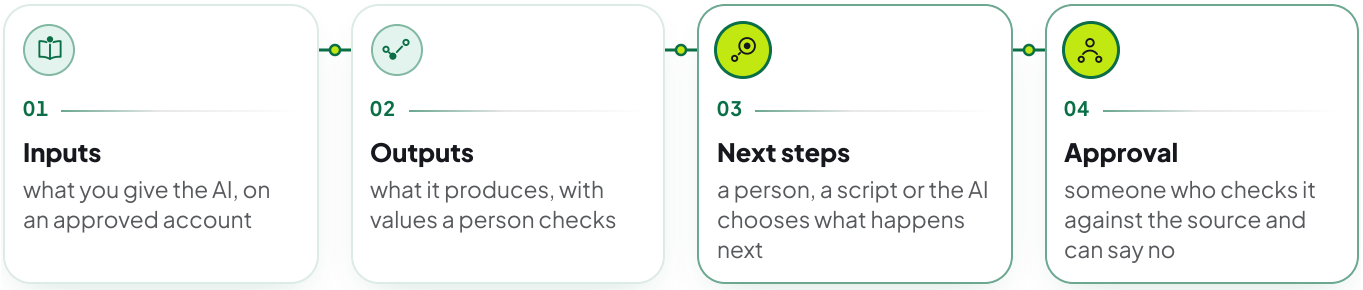
A good test fits on a receipt. Could you show the customer the reason for this message in one sentence, and would they nod? If they would wince, use less data, or tell them why they are seeing it. The same study found that trust cues softened the backlash.

9. The agency playbook

The same four checks come up across the eight chapters, in the order a piece of work moves through a team. Inputs: give the AI only the material the job needs, on an approved account. Outputs: ask for a format software can read, check the values in it, and know who owns any

image it made. Next steps: know whether a person, a fixed script or the AI decides what happens next, and where the AI has to stop. Approval: the reviewer sees the source next to the draft and has the authority to reject it. Run the checks before the team starts relying on a tool, because each one costs more to fix after something has gone wrong.

FOUR CHECKS BEFORE YOUR TEAM RELIES ON AI



A new tool or a new job starts the four checks again.

- Write the four answers into the statement of work (SOW): the approved tools and plans, the output format, which steps run on their own, and who approves the work.
- Ask every vendor the same short list. Which tier and retention apply? Is it a chatbot, a copilot, a workflow or an agent? What happens when it gets stuck? Who owns what it makes?
- Run the amnesty before the policy. Ask the team what they use, with no blame attached, then approve a few tools they will want to use.
- Test the process as well as the model. Slip a known error into a small review sample and see whether anyone catches it.

CORTEXA'S READ

None of this needs a new department. Most of it is one page of decisions, made at the start of the work, when they are still cheap to change. Teams that make them can hand AI more of the work, because they know where each process stops and who checks the results before a client sees them.

Sources

1. Cortexa Ground Truth Nº 10 — “AI, Without the Hype — The Cortexa Field Guide”
   https://www.cortexa-consulting.ai/insights/whats-real-in-ai
2. Cortexa Ground Truth Nº 20 — “Building With AI, Without the Guesswork.”
   https://www.cortexa-consulting.ai/insights/building-with-ai-without-guesswork
3. Cortexa Ground Truth Nº 30 — “Shipping AI, Without the Surprises.”
   https://www.cortexa-consulting.ai/insights/shipping-ai-without-surprises
4. Cortexa Ground Truth Nº 31 — “A context window is just a very long shelf.”
   https://www.cortexa-consulting.ai/insights/what-is-a-context-window
5. Cortexa Ground Truth Nº 32 — “Client data in ChatGPT depends on the tier.”
   https://www.cortexa-consulting.ai/insights/chatgpt-and-client-data
6. Cortexa Ground Truth Nº 33 — “Your team already uses AI. Make that use safe.”
   https://www.cortexa-consulting.ai/insights/shadow-ai-and-team-guardrails
7. Cortexa Ground Truth Nº 34 — “AI images can run in ads. Keep the receipts.”
   https://www.cortexa-consulting.ai/insights/ai-images-in-paid-campaigns
8. Cortexa Ground Truth Nº 35 — “Open-weight AI models are yours to run and to maintain.”
   https://www.cortexa-consulting.ai/insights/open-weight-ai-models
9. Cortexa Ground Truth Nº 36 — “Make the AI answer in a shape your system can read.”
   https://www.cortexa-consulting.ai/insights/structured-output-and-function-calling
10. Cortexa Ground Truth Nº 37 — “Human review of AI work has to be able to say no.”
    https://www.cortexa-consulting.ai/insights/human-in-the-loop-review
11. Cortexa Ground Truth Nº 38 — “AI personalization works until it feels like spying.”
    https://www.cortexa-consulting.ai/insights/ai-personalization-line
12. Cortexa Ground Truth Nº 39 — “Is it an agent? Ask who decides the next step.”
    https://www.cortexa-consulting.ai/insights/is-it-really-an-ai-agent
13. Anthropic — “Context windows” (Claude Docs)
    https://platform.claude.com/docs/en/build-with-claude/context-windows
14. Anthropic — “Models overview” (Claude Docs, model comparison table)
    https://platform.claude.com/docs/en/models/overview
15. Anthropic — “Effective context engineering for AI agents”
    https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents
16. Liu et al. — “Lost in the Middle: How Language Models Use Long Contexts” (preprint)
    https://arxiv.org/abs/2307.03172
17. Hsieh et al. — “RULER: What's the Real Context Size of Your Long-Context Language Models?” (preprint)
    https://arxiv.org/abs/2404.06654
18. OpenAI Help Center — “Data controls FAQ”
    https://help.openai.com/en/articles/7730893-data-controls-faq

19. OpenAI — “Data governance and compliance” (OpenAI Academy)
https://academy.openai.com/public/clubs/admins-6o6×f/resources/data-governance-and-compliance

20. OpenAI — “Data controls in the OpenAI platform” (developer docs)
https://developers.openai.com/api/docs/guides/your-data

21. Anthropic — “Is my data used for model training?”
https://privacy.claude.com/en/articles/7996868-is-my-data-used-for-model-training

22. Anthropic — “How long do you store my organization's data?”
https://privacy.claude.com/en/articles/7996866-how-long-do-you-store-my-organization-s-data

23. Microsoft and LinkedIn — “AI at Work Is Here. Now Comes the Hard Part” (2024 Work Trend Index)
https://www.microsoft.com/en-us/worklab/work-trend-index/ai-at-work-is-here-now-comes-the-hard-part — Vendor-run survey of 31,000 knowledge workers, February to March 2024; self-reported.

24. OWASP — LLM02:2026 Sensitive Information Disclosure (Top 10 for LLM Applications 2026)
https://github.com/GenAI-Security-Project/GenAI-LLM-Top10/blob/main/2026/final/LLM02_SensitiveInformationDisclosure.md

25. Stanford HAI — 2025 AI Index Report, Technical Performance
https://hai.stanford.edu/ai-index/2025-ai-index-report/technical-performance

26. Hugging Face — “Welcome GPT OSS, the new open-source model family from OpenAI”
https://huggingface.co/blog/welcome-openai-gpt-oss

27. Meta — Llama 3.1 Community License Agreement (llama-models repository)
https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/LICENSE

28. EU Artificial Intelligence Act (Regulation 2024/1689), Article 53(2)
https://eur-lex.europa.eu/eli/reg/2024/1689/oj

29. OpenAI — Structured Outputs (API guide)
https://platform.openai.com/docs/guides/structured-outputs

30. Anthropic — Tool use with Claude (Claude Platform docs)
https://docs.claude.com/en/docs/agents-and-tools/tool-use/overview

31. OWASP — LLM10:2026 Improper Output Handling (Top 10 for LLM Applications 2026)
https://github.com/GenAI-Security-Project/GenAI-LLM-Top10/blob/main/2026/final/LLM10_ImproperOutputHandling.md

32. Anthropic — “Building effective agents”
https://www.anthropic.com/engineering/building-effective-agents

33. OpenAI — “A practical guide to building agents” (PDF)
https://cdn.openai.com/business-guides-and-resources/a-practical-guide-to-building-agents.pdf

34. Microsoft — “Unlocking autonomous agent capabilities with Microsoft Copilot Studio”
https://www.microsoft.com/en-us/microsoft-copilot/blog/copilot-studio/unlocking-autonomous-agent-capabilities-with-microsoft-copilot-studio/

35. Microsoft Research — Overreliance on AI: Literature Review
https://www.microsoft.com/en-us/research/publication/overreliance-on-ai-literature-review/

36. Bansal et al. — Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance (arXiv preprint)

https://arxiv.org/abs/2006.14779

37. European Union — AI Act, Article 14 (human oversight)
https://artificialintelligenceact.eu/article/14/

38. OWASP — LLM03:2026 Excessive Agency (Top 10 for LLM Applications 2026)
https://github.com/GenAI-Security-Project/GenAI-LLM-Top10/blob/main/2026/final/LLM03_ExcessiveAgency.md

39. U.S. Copyright Office — "Copyright and Artificial Intelligence, Part 2: Copyrightability" (January 2025)
https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf

40. U.S. Court of Appeals, D.C. Circuit — Thaler v. Perlmutter, No. 23-5233 (March 18, 2025)
https://media.cadc.uscourts.gov/opinions/docs/2025/03/23-5233.pdf

41. Microsoft — "Microsoft announces new Copilot Copyright Commitment for customers"
https://blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns/

42. Google Cloud — "Protecting customers with generative AI indemnification"
https://cloud.google.com/blog/products/ai-machine-learning/protecting-customers-with-generative-ai-indemnification

43. Meta — "Expanding GenAI Transparency for Meta's Ads Products"
https://about.fb.com/news/2025/02/gen-ai-transparency-metas-ads-products/

44. Aguirre, Mahr, Grewal, de Ruyter and Wetzels — "Unraveling the personalization paradox" (Journal of Retailing, 2015)
https://doi.org/10.1016/j.jretai.2014.09.005 — Peer-reviewed. Facebook field study plus three experiments, so the pattern predates generative AI.

45. Google Ads Policy — Personalized advertising
https://support.google.com/adspolicy/answer/143465 — Vendor-stated policy; it changes, so check the live page.

46. General Data Protection Regulation — Article 21, Right to object
https://gdpr-info.eu/art-21-gdpr/ — An unofficial reproduction of the regulation's text; the official text is on EUR-Lex.