

AI OUTPUT AND THE IMPORTANCE OF HUMAN REVIEW

A practical guide to checking meaning, facts, and consequences, with 28 examples.

Client Education Series

An exclusive publication of Sandbox Technologies, Inc.

Sandbox Technologies
3868 W Carson St, Suite 316
Torrance, CA 90503
www.sandboxtech.com
Tel. 424.207.5130

SANDBOX TECHNOLOGIES

Small Business Solutions
Enterprise GrowthPath®
EGP Secure®
EGP Cloud™
ConstructIT®

AI Output and The Importance of Human Review

It's widely known that AI output, while at times highly impressive, should always be reviewed by a human, but other than generalities, or discussions of AI hallucination, memory loss when context windows become excessive, and sycophantism, reasons are not always part of the discussion. We believe that to ensure this guidance is taken very seriously, teaching our clients some of the reasons behind different behaviors will help them to not only better appreciate the risks, but also become better at spotting output issues they might otherwise have missed. Stated simply, review AI output for meaning, facts, and consequences as well as readability. The following illustrative examples show how a plausible result can depart from its source. The explanations describe possible mechanisms, not proven causes of a particular output, and the behaviors are often tuned to the behavior of some models more than others.

A practical review process

Compare the result with the source. Check facts, numbers, dates, qualifications, and the difference between discussion and approval.

Confirm that every commitment and disclosure is authorized and appropriate for the recipient. Open cited sources and check that they support the statement.

Assign review to someone who understands the work. Escalate questions beyond that person's authority before sending, relying on the result, or allowing a system to change records.

Check the action as well as the wording. A correct draft does not authorize sending it, changing files, or carrying out a transaction.

SANDBOX TECHNOLOGIES

Copyright 2026, Sandbox Technologies, Inc.

1. Risk - Reassurance in an original message or prompt becomes re-written as a promise

Issue	An investigation becomes a guaranteed remedy.
What to watch out for	Does the message make an unintended commitment?
Why AI may do this	Familiar service language may tend to favor a satisfying resolution over the original limits.
Example	Source Document: “We will investigate and discuss next steps.” AI Re-Write: “We will resolve this at no cost to you.”

2. Risk - An acknowledgment becomes an admission of fault

Issue	Neutral wording becomes an unsupported statement accepting responsibility.
What to watch out for	Does the response admit a failure, cause, or responsibility that has not been established?
Why AI may do this	A request for an empathetic apology may produce familiar blame-accepting language, adding facts or causation absent from the source.
Example	Source Document: “We received an IRS notice stating that your tax filing was received late.” AI Re-Write: “We apologize for failing to submit your return on time.” Potential consequence: The rewrite accepts responsibility that the original notice did not establish.

3. Risk - An estimate becomes a fact

Issue	An expected date becomes a confirmed date.
What to watch out for	Has meaningful uncertainty disappeared?
Why AI may do this	Making wording more direct can remove qualifications that carry meaning.
Example	Source Document: “We expect delivery Friday, subject to carrier confirmation.” AI Re-Write: “Your order will arrive Friday.”

SANDBOX TECHNOLOGIES

Copyright 2026, Sandbox Technologies, Inc.

4. Risk - A condition disappears

Issue	A conditional benefit becomes unconditional.
What to watch out for	Are conditions and exceptions preserved?
Why AI may do this	Shortening text may discard a condition along with unnecessary words.
Example	Source Document: “The upgrade is included if you renew for another year.” AI Re-Write: “Your upgrade is included.”

5. Risk - A source’s meaning gets reversed

Issue	A restriction becomes permission to proceed.
What to watch out for	Does the summary preserve the actual instruction?
Why AI may do this	Compression may lose the relationship between an action, restriction, and timing.
Example	Source Document: “Do not deploy until testing is complete.” AI Re-Write: “Proceed with deployment while testing continues.”

6. Risk - Missing information gets invented

Issue	An unsupported detail fills a gap.
What to watch out for	Which details are supported by the source?
Why AI may do this	A familiar document pattern may be completed with a plausible detail.
Example	Source Document: “The schedule will be agreed after approval.” AI Re-Write: “Installation will be completed by November 15.”

7. Risk - A citation supports nothing

Issue	A reference is nonexistent or does not support the claim.
What to watch out for	Does the source exist and support the statement?
Why AI may do this	Generating a convincing reference differs from retrieving and checking a source.
Example	Source materials contain no study. AI Re-Write: “A 2025 industry study found a 37% productivity gain,” with an unverified reference.

SANDBOX TECHNOLOGIES

Copyright 2026, Sandbox Technologies, Inc.

8. Risk - Numbers produce a wrong conclusion

Issue	Correct inputs lead to an incorrect percentage.
What to watch out for	Are calculations, units, and comparisons correct?
Why AI may do this	Generating numerical language does not ensure the calculation was performed correctly.
Example	Source Document: “Sales rose from $100,000 to $120,000.” AI Re-Write: “Sales increased 25%.” Correct increase: 20%.

9. Risk - A narrow finding becomes a broad claim

Issue	A limited observation becomes a general conclusion.
What to watch out for	Are scope, sample size, and qualifications retained?
Why AI may do this	A summary may retain a headline while dropping its limits.
Example	Source Document: “Three of 40 trial participants reported slow loading.” AI Re-Write: “Customers are dissatisfied with performance.”

10. Risk - Private details reach the wrong audience

Issue	Internal information appears in an external message.
What to watch out for	Is every detail appropriate for the recipient?
Why AI may do this	Available context may be treated as relevant without recognizing audience boundaries.
Example	Internal note: “Our margin on this project is 35%.” AI Re-Write: “The price includes our 35% margin.”

11. Risk - Discussion becomes a decision

Issue	An option under discussion becomes an approved action.
What to watch out for	Does the output distinguish suggestions from approvals?
Why AI may do this	A summary may turn an unresolved conversation into a more definite account.
Example	Source Document: “We discussed reducing staffing; no decision was made.” AI Re-Write: “Management agreed to reduce staffing.”

SANDBOX TECHNOLOGIES

Copyright 2026, Sandbox Technologies, Inc.

12. Risk - Capabilities get overstated

Issue	An unavailable feature appears in a sales description.
What to watch out for	Does the description match the actual offering?
Why AI may do this	Knowledge of similar products can fill gaps in company-specific information.
Example	Source Document: “Our product exports CSV files.” AI Re-Write: “Our product synchronizes automatically with your accounting system.”

13. Risk - A procedure omits a critical step

Issue	A plausible procedure leaves out a safeguard.
What to watch out for	Are steps complete and appropriate for the actual system?
Why AI may do this	A general procedure may omit situation-specific dependencies or safeguards.
Example	Required process: Preserve a recoverable copy, validate the import, then replace records. AI Re-Write: Import the new file and replace existing records.

14. Risk - Tone changes the relationship

Issue	A firm message becomes accusatory.
What to watch out for	Does the tone fit the circumstances and relationship?
Why AI may do this	A request to make writing firm or friendly can exaggerate that quality.
Example	Source Document: “Our records show the invoice remains unpaid. Please confirm its status.” AI Re-Write: “You have ignored your payment obligations.”

15. Risk - Responsibility gets assigned to the wrong person

Issue	A statement or task is attributed to someone else.
What to watch out for	Verify who said, approved, owns, and must perform each action.
Why AI may do this	A summary may lose speaker labels or merge nearby statements.
Example	Source Document: “Maya will obtain the quote. Luis will approve the purchase.” AI Re-Write: “Luis will obtain the quote; Maya will approve it.”

Copyright 2026, Sandbox Technologies, Inc.

16. Risk - An option becomes a requirement

Issue	Optional wording becomes a mandatory instruction.
What to watch out for	Check may, should, must, and the authority behind each requirement.
Why AI may do this	A rewrite may strengthen language while trying to make instructions decisive.
Example	Source Document: “Staff may attend the optional Friday workshop.” AI announcement: “All staff must attend Friday’s workshop.”

17. Risk - Ambiguity gets silently resolved

Issue	An unclear reference becomes a definite interpretation.
What to watch out for	Identify ambiguous names, pronouns, dates, and instructions; ask rather than assume.
Why AI may do this	A fluent answer may select one plausible interpretation without flagging alternatives.
Example	Source Document: “After Jordan meets Casey, send them the revised quote.” AI task list: “Send Casey the revised quote,” although the recipient was never confirmed.

18. Risk - Events appear in the wrong order

Issue	The sequence of events changes.
What to watch out for	Compare event dates and before or after relationships with the record.
Why AI may do this	Compressing several updates can disconnect an event from its time.
Example	Source Document: “Service failed Monday. The update was installed Tuesday.” AI incident summary: “Service failed after Tuesday’s update.”

19. Risk - Units or reporting periods get mixed

Issue	A number keeps its digits but changes its basis.
What to watch out for	Check currencies, units, recurring periods, and whether amounts are totals or per person.
Why AI may do this	Related figures may be combined without preserving their labels.
Example	Source Document: “The license costs $80 per user per month.” AI budget: “The annual license cost is $80 per user.”

Copyright 2026, Sandbox Technologies, Inc.

20. Risk - Old information overrides a newer correction

Issue	A superseded fact is presented as current.
What to watch out for	Check source dates, versions, and explicit corrections.
Why AI may do this	Earlier material or familiar information may outweigh a later update.
Example	Source Document: “Original delivery: June 3. Updated confirmation: June 10.” AI client message: “Delivery is confirmed for June 3.”

21. Risk - Correlation becomes causation

Issue	Two events occurring together become an explanation of cause.
What to watch out for	Look for evidence of causation, competing explanations, and limits of the analysis.
Why AI may do this	A narrative can supply a causal connection the data does not establish.
Example	Source Document: “Sales rose during the month the new campaign launched.” AI Re-Write: “The campaign caused the increase in sales.”

22. Risk - Important evidence gets overlooked

Issue	A long record is summarized without a material fact.
What to watch out for	Check the result against a list of required facts and contrary evidence.
Why AI may do this	Models may use long inputs unevenly and miss relevant material within them.
Example	Source Document: A long project file includes an unresolved accessibility defect. AI readiness report: “No outstanding launch issues,” omitting the defect.

23. Risk - Table values attach to the wrong labels

Issue	Extracted data shifts between rows or columns.
What to watch out for	Match each value to the original row, column, heading, and footnote.
Why AI may do this	Layout conversion can separate a figure from the label that gives it meaning.
Example	Source table: North revenue $40,000; South revenue $70,000. AI Re-Write: “North generated $70,000; South generated $40,000.”

Copyright 2026, Sandbox Technologies, Inc.

24. Risk - Missing data becomes zero

Issue	Unknown or unavailable values are treated as measured results.
What to watch out for	Distinguish zero, blank, not applicable, and not yet reported.
Why AI may do this	A calculation or summary may fill empty fields without checking their meaning.
Example	Source Document: Three branches reported sales; the fourth has not submitted data. AI Re-Write: “The fourth branch recorded no sales.”

25. Risk - An answer agrees with a false premise

Issue	A leading question receives agreement instead of scrutiny.
What to watch out for	Test the premise and ask for evidence that would contradict the conclusion.
Why AI may do this	Some models favor agreement with the user’s stated view over correcting it.
Example	Prompt: “Since all complaints are from new customers, explain why onboarding failed.” AI accepts the premise even though the complaint log includes long-standing customers.

26. Risk - Completion is claimed without evidence

Issue	A draft or attempted action is reported as completed work.
What to watch out for	Verify the actual saved file, sent message, system record, or independent confirmation.
Why AI may do this	Generated status language may describe the intended outcome without checking the action result.
Example	AI says: “The customer record has been updated.” The system shows the save failed and the original record remains unchanged.

27. Risk - Recommendations reflect an unsupported stereotype

Issue	An evaluation uses assumptions unrelated to the actual evidence.
What to watch out for	Compare recommendations with consistent, job-relevant criteria and the supporting record.
Why AI may do this	Patterns learned from examples can reproduce biased associations.
Example	Source Document: Two applicants have comparable documented qualifications. AI favors one based on an assumption about availability drawn from personal background.

Copyright 2026, Sandbox Technologies, Inc.

28. Risk - The output contradicts itself

Issue	Different sections give incompatible facts or instructions.
What to watch out for	Compare the summary, detail, tables, and final recommendations for consistency.
Why AI may do this	Separate passages may be generated without reconciling earlier statements.
Example	AI Re-Write: The summary promises a 12-month term; the pricing section specifies a 24-month minimum.

Research context

These illustrative business scenarios are not reported incidents or a frequency ranking. The research below supports broad concerns; it does not establish the cause of any individual example.

NIST AI 600-1 Generative Artificial Intelligence Profile

https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

Anthropic Towards understanding sycophancy in language models

https://www.anthropic.com/research/towards-understanding-sycophancy-in-language-models

Liu and colleagues Lost in the Middle

https://arxiv.org/abs/2307.03172

We're here to help and have structured services to address each of these areas individually or as part of a coordinated plan. To discuss what deserves attention first, contact your Sandbox Technologies Account Manager or Consulting CIO.

These publications provide general information. Review our Website Disclaimer and assess recommendations against your organization's circumstances.

Copyright 2026, Sandbox Technologies, Inc.