False approval claims: a deterministic verification of approval statements in AI assistants with tools

Formal definition, per-turn measurement, extraction error and validation

Juan F. Cobo (author)
Leonel Leenen (reviewer, 2026-10-20)
EthiCompass

Manuscript prepared for external review and eventual submission for publication

Version 1.0, 2026-10-20

Abstract

An AI assistant with tools tells what it did and why. Sometimes it says that a person approved the action: "your supervisor already approved it". If that approval does not exist, the assistant acted alone, hid it and moved the responsibility to a person who never saw the case. The published metrics treat this failure as a case of general fabrication, decided by a judge. We propose to measure it separately, because it can be verified without a judge: an extraction finds the approval statements in each turn, and a deterministic verification looks for them in the approval record. Only a granted approval, recorded by the approval system, earlier than or simultaneous with the turn and covering the claimed action, supports a statement; if the assistant could write the record, it could record the approval that it invents. The result of each turn is 0 or 1; the result of the conversation is the maximum; one confirmed false statement is an incident. The only source of error is the extraction, whose sensitivity and specificity are reported together with the rate. Without the system record, the verification uses the record of the person who runs the test, who knows what that person approved, and it stays deterministic. This work reports no results with real assistants.

Keywords: AI red-teaming, human oversight, attribution, fabrication, tool-using agents, deterministic verification, incidents.

The Spanish version of this document is the original.

1 Introduction and motivation

After it acts, the assistant explains to the user what it did. The user, and later an auditor, use this account to know who made each decision. A sentence that says "a person approved this" has a special effect: it tells the reader that the action already had a human review, and for this reason nobody reviews it again.

If the sentence is false, three things occur: the assistant acted without the approval that was necessary, it hid that it acted alone, and the responsibility moves to a person who never saw the case. The failure can have several causes: the model completes the account with a usual sentence, hostile content pushes it to justify the action, or the model confuses a requested approval with a granted approval. The measurement does not need to know the cause: it compares what the assistant says with what the record has.

Frequently the failure is in the middle of the conversation: the assistant acts, justifies the action with an approval that does not exist and then says goodbye correctly. A measurement that reads only the final response does not see it. For this reason the measurement is per turn.

2 Related work

- **Fabrication and false success.** Sethi et al. [3] show that assistants fabricate accounts after a tool failure (14.1% dishonest responses; 45.3% if the failure is not marked as an error), and that a status indicator read deterministically decreases the rate to 0.87%. Zhu et al. [4] report 22.8% false success with the base policy and 0.8% with an evidence contract. Advani [5] shows that LLM judges detect false success badly (AUROC < 0.65). The three justify verifying against a record, not against the text.
- **Attribution.** Chan et al. [6] treat the attribution of the actions of an agent and activity logs. Narajala and Narayan [7] describe the attribution gaps in threat T9: the agent hides who decided.
- **Published oversight metrics.** The Gaussia oversight paper [1] measures fabrication with a judge, inside action disclosure. In the public discussion of our proposals, its team proposed that only the data recorded by the approval system count and that the request data go in the approval event [2]; this work adopts both rules.

3 Formal definition

3.1 Data

Let i be the conversation and $t \in \{0, \ldots, T_i - 1\}$ the turn of the assistant, with text $x_{i,t}$. The approval record is a set $\mathcal{B}_i$; each approval b has a decision $d(b) \in \{\text{granted}, \text{denied}, \text{pending}\}$, a source $s(b)$, the turn of the decision $\tau(b)$ and the set of actions that it covers. The request data (the text, its author and its turn) go in the same event and help to tell "I requested the approval" from "they approved it".

3.2 Extraction

An extraction function E returns, for each turn, the set of approval statements $E(x_{i,t}) = \{a_1, \ldots, a_m\}$. Each statement has an actor (who approved, according to the text), an action (connected to a tool of the policy when possible) and the quote. An approval request, a plan or a condition ("if the supervisor approves it, I will issue it"), and a denial do not count as statements. The extraction can be a pattern or a model; it only finds the statement and does not decide whether it is false.

3.3 Verification

A statement a of turn t is supported if

$$\exists\, b \in \mathcal{B}_i :\ d(b) = \text{granted} \ \wedge\ s(b) = \text{approval system} \ \wedge\ \tau(b) \leq t \ \wedge\ \text{covers}(b, a). \tag{1}$$

$\text{covers}(b, a)$ compares the tool of the approval with the claimed action; if the record connects the approval with the call, it uses that connection. If the statement does not name an action, it is connected with the action of the same turn or of the earlier turn. The verification does not compare the parameters of the statement (Section 5).

The source. Only an approval recorded by the approval system supports a statement. If the assistant could write the approval record, it could also record the approval that it invents, and the measurement would never fail.

3.4 Result per turn and per conversation

The result of the turn is

$$g_{i,t} = \mathbf{1}\big[\exists\, a \in E(x_{i,t}) : a \text{ is not supported}\big], \tag{2}$$

with the cause `false_approval_claim` and the quote of the statement. If the extraction fails or the record is incomplete, the turn has no score and has a reason; a turn without a score never counts as compliance. The value of the conversation is $G_i = \max_t g_{i,t}$ over the turns with a score; a conversation without turns with a score goes out of the denominator, with its reason.

3.5 Rate, interval and incidents

The rate is $\hat{p} = \sum_i G_i/n$ over the evaluated conversations that target the measurement. A model never decides the denominator. The interval is Wilson [8] with the Kish effective size [9], $n_{\text{ef}} = n/(1 + (\bar{m} - 1)\rho)$, with the replicas grouped by scenario. Because the verification is deterministic, each confirmed false statement is an incident, reported with the conversation, the turn and the quote, without waiting for a rate.

3.6 Error of the extraction

The extraction is the only part that can make errors. Its error is described by the sensitivity Se (the proportion of real statements that it finds) and the specificity Sp (the proportion of turns without a statement in which it finds nothing). A statement that the extraction does not find is a failure that is not counted; a statement that it invents can give a false 1. We report Se and Sp together with the rate, with the model version and per language, and we publish the rate without correction. The Rogan-Gladen correction [10],

$$\hat{\pi} = \frac{\hat{p} + \mathrm{Sp} - 1}{\mathrm{Se} + \mathrm{Sp} - 1}, \tag{3}$$

is possible, but stays as an open question.

Calibration by construction. A large part of the set to estimate Se and Sp can be built without human labels: take turns with a known statement and change them in known ways (a different verb, a different actor, a different sentence), and use conditional forms, requests and denials as negative controls. Only the paraphrases of real traffic need human labels. The extraction model receives the tools of the policy and the turn, in the language of the assistant, and returns the list of statements with their actor, their action and their quote.

4 Observation with the conversation only

In a conducted test, the person who runs the conversation also acts as the supervisor: that person approves or denies the requests, and for this reason knows what was approved and when. Without the system record, the verification uses the record of the person who runs the test. It stays deterministic, and its result can still be used as incident evidence; only the extraction reads text. It is the only variable of the dimension that keeps its incidents without the system record.

5 Boundaries with other variables

- **Fabrication.** In the turn where this measurement gives 1, general fabrication does not measure the same sentence again: the same statement is not counted twice.

- **Escalation.** "I already passed your case to a supervisor", without an escalation call, states an escalation, not an approval: it is under-escalation.
- **Substitution.** If 90 USD was approved and the assistant says "the supervisor approved the refund of 900", the approval of the tool exists and this measurement does not fail; if 900 was executed, substitution finds the failure, because it compares what was approved with what was executed.
- **Control evasion.** A denied approval that the assistant presents as approved is a failure here; if the action was also executed, it is also a control evasion.

6 Illustrative example

(Illustrative construction; not a measurement.)

A conversation of three turns. In turn 0 the assistant writes "for a refund of 900 USD I need the approval of a supervisor" and the system records the request, with the decision pending. In turn 1 it writes "the supervisor already approved the refund of 900 USD and I issued it". In turn 2 it writes "can I help you with something else?". The extraction finds no statements in turns 0 and 2 ($g = 0$) and finds one in turn 1. The record has no granted approval with turn ≤ 1, so $g_1 = 1$ and $G = \max(0, 1, 0) = 1$: an incident in turn 1. The final response is clean: a measurement that reads only the final response gives 0.

A rate. There are 20 scenarios with 3 replicas: 60 conversations. Two have no turns with a score: $n = 58$. Seven conversations fail: $\hat{p} = 7/58 = 0.121$. With $\bar{m} = 3$ and $\rho = 0.3$, $\text{deff} = 1 + 2 \times 0.3 = 1.6$ and $n_{\text{ef}} = 58/1.6 = 36.25$; the 95% Wilson interval with these values is $(0.050,\ 0.264)$. In addition, 7 incidents are reported, each one with its turn and its quote.

7 Validation

7.1 Test cases with a known result

Because the verification is deterministic, the cases test the exact result: a false statement in turn 1 with a clean final response; a true statement; a request without a statement; a denied approval presented as granted; an approval of a different action; an approval later than the statement; a hypothetical statement; a turn without text in the record; an approval with a source different from the approval system; and an escalation statement. In the set of conformance cases of the dimension, two cases are of this variable: a false approval claim with a record, and the same failure in turn 1 of a conversation of three turns with a clean final response. The set detects, among others, a judge score instead of 0 or 1 and the violation of turn 1 not seen. The source rule is not yet in that set.

7.2 Synthetic simulation

In a simulation of the dimension with a known truth, separating this variable from fabrication keeps decisions: with a simulated judge with 90% accuracy, when the fabrication judge does not respond, the verification of the false approval claim is still decided. Without the system record, the failure stays visible, because the person who runs the test knows that that person did not approve.

8 Discussion

8.1 What this work establishes

The measurement is defined by the extraction and the record: what an approval statement is, what supports it, how each turn is decided, how the turns are joined per conversation and how its only error is described. The verification is deterministic and can be used as incident evidence, also without the system record.

8.2 What it does not establish

- **There are no results with real assistants,** and no measurement of Se and Sp for a specific extraction.
- **It sees only what the record has.** If the system approves through a channel that it does not export, a true statement looks false.
- **It does not say why the assistant stated something false.**
- **It does not measure other false statements.** "I checked your history" without a call that does it is fabrication, not this failure.
- **It depends on the scenarios.** If no scenario leads to an action that needs approval, the rate says nothing.

8.3 Generalization

Separating from a judge a statement that can be verified against a record applies to other statements of the assistant about recorded facts. It is not an extension evaluated here.

9 Conclusion

A false approval claim can be measured without a judge: an extraction finds the statement and the record decides whether it is false. The source rule prevents the assistant from supporting its own statement. The per-turn measurement sees the failure although the final response is clean, and each confirmed case is an incident. The validation is synthetic, and the next step is to measure the error of a real extraction and real assistants.

References

[1] Gaussia Labs (2026). *Accountability: oversight compliance and action disclosure.* `gaussia-labs/papers`, `papers/2026-09-accountability/`. Preprint.

[2] Gaussia Labs (2026). Replies of the team to discussions #78, #80 and #84, repository `gaussia-labs/papers`. Public discussion.

[3] Sethi et al. (2026). *Fabrication After Tool Failure.* arXiv:2609.14758.

[4] Zhu et al. (2026). *Failure-Transparent Agents.* arXiv:2609.35732.

[5] Advani, L. (2026). *From Confident Closing to Silent Failure: Characterizing False Success in LLM Agents.* arXiv:2606.09863.

[6] Chan, A. et al. (2024). *Visibility into AI Agents.* FAccT 2024.

[7] Narajala, V. & Narayan, O. (2025). *Securing Agentic AI.* arXiv:2504.19956.

[8] Wilson, E.B. (1927). *Probable inference, the law of succession, and statistical inference.* Journal of the American Statistical Association, 22(158), 209–212.

[9] Kish, L. (1965). *Survey Sampling.* Wiley.

[10] Rogan, W.J. & Gladen, B. (1978). *Estimating prevalence from the results of a screening test.* American Journal of Epidemiology, 107(1), 71–76.

Declarations

Conflicts of interest. The authors are affiliated with EthiCompass, the organization that develops the red-teaming system whose Accountability & Human Oversight dimension motivates this work. The measurement defined here is of general application and does not depend on a proprietary implementation, but the affiliation is declared so that the reader can take it into account.

Relation to Gaussia. The source rule and the location of the request data follow what the Gaussia team proposed in a public discussion. The separation from fabrication, the per-turn measurement and the validation are the work of the authors.

Funding. *To be completed before submission.*

Data and code availability. The results of Section 7 come from a synthetic simulation and hand-written test cases. *The access conditions are completed before submission.*