

Peer Review

Review of: "WeAudit: Scaffolding User Auditors and AI Practitioners in Auditing Generative AI"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper introduces WeAudit, a workflow and a web prototype meant to help ordinary users audit text-to-image systems and to help industry practitioners act on what those users find. The design rests on a formative study: think-alouds with 11 end users recruited through Craigslist and Nextdoor (§3.1.1) and semi-structured interviews with 7 industry practitioners (§3.1.2). From those, the authors derive six design goals, DG1 through DG6 (§3.2), and map each to a piece of the workflow in Table 1. The system instantiates them as seven features (§4.2): pairwise comparison, a prompt history sidebar, a worked example repository of 50 curated examples, social augmentation showing what tags other auditors have used, a structured report portal, a Discourse-based forum, and a crowd verification survey. The backend runs Stable Diffusion 2.0 through the Replicate API on Amazon Lightsail (§4.2.8).

The evaluation has two halves. Forty-five students at a US university used WeAudit in one of two synchronous sessions (§5.1, N1=15, N2=30), each session consisting of a 10-minute onboarding, 40 minutes of auditing, 10 minutes of survey, and a 30-minute group discussion, after which the forum stayed open for three weeks of voluntary use. That produced 164 audit reports, 372 tags, 62 forum comments from 17 users across 43 reports, 45 post-audit surveys, and 2,249 logged interactions (§5.3). Eighty-four Prolific crowdworkers then verified the reports. Ten industry practitioners were interviewed with those reports and verification results in hand (§5.2). The claimed findings, set out in §1 as four contributions, are that WeAudit helps users notice harms they would otherwise miss, deepens and broadens their audits, produces reports practitioners find actionable, and raises users' awareness of AI harms.

Let me say up front what I think is genuinely good here, because a lot of it is. The formative-to-design-goal-to-feature chain is unusually explicit for this literature, and Table 1 makes it auditable rather than asserted in prose. The decision in §5.1.1 to run the first study on students specifically to avoid burdening marginalized communities is the right call and is argued, not just announced. §7 names the student sample and the US-and-large-company practitioner skew as limitations before a reviewer has to. Table 4 reconciles four ways: majors sum to 45, race to 45, gender to 45, program to 45, and the tag counts in §6.2.3 (205 affected-group tags plus 167 harm-type tags) sum to the 372 reported in §6.3. I checked the mean age in Table 3 by hand: the eleven ages sum to 522, and $522/11 = 47.4545$, so the "47.45 average" in §3.1.1 is right. Small thing. It tells me someone was careful.

My problems are with the inferential language draped over descriptive data, with several numbers that do not survive being added up, and with one statistic in §6.3.2 that I believe cannot be what the paper says it is.

2. Whether the conclusion holds on the data

2.1 There is no control condition anywhere, and the verbs do not reflect that. Every participant used the full WeAudit system. No one audited Stable Diffusion 2.0 with the comparison feature disabled, or with the worked examples hidden, or with a free-text report box instead of the structured portal. That is a defensible choice for a first deployment study, and §5.1.1 half-acknowledges it by framing the goal as "an initial understanding of WeAudit's strengths and limitations." But the contribution list in §1 says WeAudit "enhances the depth and breadth of user audits," and the §6.2 heading says worked examples and social augmentation "enhanc[e] the Depth and Breadth of Audits." Enhance relative to what? Depth and breadth are comparative words with no comparator in this design. The same applies to the abstract's claim that WeAudit supports users "in articulating their findings in ways that industry practitioners can act upon": the ten practitioners in §5.2 read reports produced through the structured portal and never read reports produced without it, so their enthusiasm speaks to the reports' absolute quality, not to the portal's contribution.

2.2 The comparison-feature evidence is weaker than it looks. §6.1 reports that auditors spent on average 82.6% of their auditing time on the comparison page and that 74% of the 164 reports involved comparisons of different prompts. The second number is offered as support for the first claim, but read together they cut the other way. If users spent 82.6% of their time in comparison mode and only 74% of reports came out of it, then per unit of time the comparison view was slightly less productive of reports

than the single-prompt view. I do not want to over-read that either, since 74% of 164 is 121.4 reports, so the underlying count is 121 or 122, and the gap is not large. The point is that the numbers as presented are compatible with the comparison page simply being where the interface puts people, which is what §4.2.1 describes: users land on a single-prompt view and toggle. F17's Pakistani-versus-Indian example is a good qualitative illustration, and I would keep it. The two percentages do not add quantitative weight to it.

2.3 The correlation in §6.2.1 is the only inferential statistic in the paper, and it is underspecified. The text reports "a positive relationship ($r=0.769$, $p<0.001$) between the number of unique worked examples a user viewed and their rate of discovering and reporting harmful AI behaviors (per the number of prompts they explored)." I could not find the sample size, the correlation type (Pearson or Spearman), a confidence interval, or any treatment of the two sessions as a grouping factor. Assuming $n = 45$, $r = 0.769$ is a very large effect for behavioral log data, which makes me want to see the scatter; Figure 8 is captioned but, as rendered on Qeios, carries no legible axis values or point cloud, so I could not inspect it. There is also a structural worry about the dependent variable. Reports per prompt explored puts prompts in the denominator, and time in the session was fixed at 40 minutes, so a user who spent that time reading examples necessarily explored fewer prompts, which inflates the ratio independently of any effect on noticing. Engagement is the obvious confound on the other side: a motivated participant clicks more of everything. The paper's own next sentence concedes that heavy example-viewers "often submitted direct variations on worked examples," which is evidence that example exposure shaped *what* was reported. That is a different and more interesting finding than the correlation.

2.4 The verification statistic in §6.3.2 is internally impossible as written. This is the item I would most want the authors to check. The text reports that verifiers agreed they could understand the reporter's harm rationale with "an average of 80.35% agreement ($\sigma=1.64\%$) across all reports." If that σ is the standard deviation of the per-report agreement rate, then essentially every one of the 164 reports sat between roughly 77% and 84% agreement. Two sentences later, the paper discusses "reports with low agreement across verifiers," and a paragraph after that, it describes cases where "a majority of verifiers agreed that they could not understand why the user found the AI output harmful." A majority not understanding means agreement below 50%, which is more than eighteen standard deviations below the stated mean. Those two passages cannot both be true of the same distribution. Either σ is a standard error, or it describes variation across something other than reports, or one of the numbers is wrong.

Because the marginalized-voices argument in §6.3.2 and the design implication in §8.3 both rest on the low-agreement cases existing, this is load-bearing.

2.5 The verifier sample does not add up. §5.1.1 states: "Among the 84 crowdworkers, 55 identified as White, 19 as Black, 11 as Asian, 14 as mixed, and 15 as other." Summing: $55 + 19 + 11 + 14 + 15 = 114$, which is 30 more than 84. Multi-select would explain it, but the sentence does not say so, and the mixed category makes overlap ambiguous. The same paragraph gives a male-to-female ratio of about 2:1 and a mean age of 36.04. Note the contrast with the auditors in Table 4, who were 30 female, 13 male, 2 non-binary, and 31 of 45 Asian. So the group producing the reports and the group judging them differed sharply on both gender and race. The paper hypothesizes in §6.3.2 that unexplained disagreement "may have been cases where the user auditor came from a marginalized group that was underrepresented among verifiers," and then leaves it as a hypothesis. The data to test it already exist: verifier demographics were collected, and the report portal has the checkbox "This audit report is relevant to my own identity" (§4.2.5). Cross-tabulate them.

2.6 Two counts of the follow-up survey disagree. §5.1.1 says 37 participants were contacted and "17 participants responded to the follow-up survey." §5.3 lists "19 follow up survey responses" among the data collected. §6.6 then works from 17, of whom 15 reported changed awareness. Seventeen or nineteen, one of these is wrong, and since contribution 4 in §1 rests on this survey, the denominator matters.

2.7 The awareness finding is a leading question answered by a self-selected third of the cohort. §5.1 quotes the item: "Did the session on auditing text-to-image models change your awareness of ways generative AI might perpetuate biases or cause harm?" That presupposes the direction of the answer and invites agreement from people who volunteered to be re-contacted four months after taking part in a study about AI harm. Fifteen of 17 said yes, and 17 is 38% of the 45 who took part. There is no pre-session measure of awareness and no group that did not audit. The qualitative material from F45, F07, and F44 is worth reporting as what participants say the experience did for them. Contribution 4 in §1, which asserts that WeAudit "helps users increase their awareness and understanding of AI harms," asserts more than that.

2.8 "Three-week user study" oversells the exposure. The abstract and §5 both describe a three-week study. What §5.1 actually describes is a single 90-minute synchronous session followed by three weeks in which continued forum use was voluntary, and §6.4 reports the uptake: 17 users, 62 comments, 43 reports touched. So 28 of 45 participants contributed nothing during the three weeks. The honest description is a

90-minute session with a three-week optional follow-on, and the 17/45 participation rate is itself a finding about sustaining engagement that belongs in §8.1, where sustaining participation is the topic.

2.9 A session-order confound goes unexamined. Social augmentation shows users what other auditors have already tagged (§4.2.4). The two sessions were run separately with N1=15 and N2=30 (§5.1). Whichever ran second inherited a populated tag distribution and a forum with content; whichever ran first did not. All of §6.2.2's claims about social augmentation nudging users toward under-explored topics, including the 12 of 45 who said the tag distribution influenced them, are therefore potentially session-dependent. The paper never splits any result by session.

2.10 Small inconsistencies in the participant reporting. §3.1.1 says the 11 formative participants held "2 high school degrees, 2 Associate degrees, 5 bachelor degrees, 1 master degree, and 1 Ph.D." Table 3 lists six bachelor's degrees (U01, U03, U04, U07, U08, U11), one high school or lower (U02), one associate's (U09), one "Some college, no degree" (U06), one master's (U10), and one doctorate (U05). The prose and the table do not match, and the "some college" category is missing from the prose entirely. Separately, Table 5's caption says "All but one industry practitioner (P13, start up company) participants are from large technology companies," while §7 says the sample had "only one participant from the U.K. and two from startups." One startup or two.

3. Reproducibility

This is a systems-and-interviews paper rather than a model paper, so the reproducibility bar is different, but several items that would let another group rebuild this study are absent.

Generation parameters. §4.2.8 names Stable Diffusion 2.0 via Replicate and nothing else. No checkpoint hash, no sampler, no inference step count, no guidance scale, no negative prompt, no output resolution, no seed policy, no statement about whether Replicate's safety filter was on or off. For a paper whose entire subject is the distribution of harmful outputs, the safety filter setting alone can change what participants saw. Nor does §4.2 say how many images each prompt produced in WeAudit; the formative study used 8 images through Dream Studio (§3.1.1), and the only clue about the system itself is a participant quote in Table 2 mentioning "all the 6 generated pictures." A reader should not have to infer an interface parameter from a quotation.

Materials. The repository of 50 worked examples (§4.2.3), the harm taxonomy mapped from Shelby et al. (§3.1.1), the tag vocabularies for "types of harms" and "affected groups," the verification survey instrument (§4.2.7), and the onboarding deck are all described but not reproduced. §5.1 says the post-

audit and follow-up surveys are in supplementary materials; I could not locate supplementary materials attached to the version on Qeios. The example-diversity heuristic in §4.2.3 is stated as a rule (deprioritize tags the user has already used) but never given as pseudocode, and it is never evaluated, even though §6.2.3 shows tags still concentrating hard on race (76 of 205 affected-group tags) and stereotyping (93 of 167 harm-type tags).

System availability. Footnote 4 removes the deployment link for double-blind review, and §4.2.8 repeats "link redacted for review." On a published preprint with named authors, this should be restored, along with a code release. §4.2 says the team plans to host WeAudit publicly; no repository is cited.

Qualitative analysis procedure. §5.3 says two authors open-coded all reports and discussions and that the interview data were analyzed with reflexive thematic analysis. For reflexive TA, declining to report inter-rater reliability is principled, and I would not ask for a kappa. But the coding also produced counts that appear as evidence, for instance, the 74% of reports involving comparisons (§6.1) and the "dozens" of hard-to-interpret reports (§6.3.1). Counts need a codebook and a stated rule for what qualifies, and neither is given.

Verification design. How many of the 84 crowdworkers rated each report? Were assignments random or balanced? Was a report seen by the same verifiers for all four criteria (clarity, harmfulness, relevance, reasonableness)? §4.2.7 and §6.3.2 answer none of these, and the σ question in §2.4 above cannot be resolved without them. The authors cite prior work of their own on a crowdsourcing pipeline for verifying user audit reports; pointing to that citation for the design parameters would cost one sentence.

Compensation. Not reported for anyone: not for the 11 formative participants, not for the 45 student auditors, not for the 84 Prolific verifiers, not for the practitioners. §6.5 is titled around invisible labor and fair compensation, and §8.1 argues that companies should pay user auditors properly. A paper making that argument should state its own rates.

4. What specifically to fix

1. §6.3.2, the sentence giving "an average of 80.35% agreement ($\sigma=1.64\%$) across all reports." Recheck this. As written, it is incompatible with the low-agreement cases discussed in the same subsection. State what the dispersion is computed over, give the per-report distribution (a histogram would do), and say how many verifiers rated each report.
2. §5.1.1: the crowdworker race breakdown sums to 114 across 84 people (55 + 19 + 11 + 14 + 15). If the item allowed multiple selections, say so; otherwise, correct the counts.

3. Reconcile §5.1.1 ("17 participants responded") with §5.3 ("19 follow-up survey responses"), and make §6.6 use whichever is right.
4. Rewrite the comparative verbs. The §1 contribution list ("enhances the depth and breadth"), the §6.2 heading, and the abstract sentence about articulating findings actionably should be restated descriptively, since nothing in the design licenses a comparison. An alternative, which I would find more persuasive than rewording, is to run one small arm: the same task with the structured portal replaced by a free-text box, and have practitioners rate both sets blind.
5. §6.2.1, the $r=0.769$ result needs n , the correlation type, a confidence interval, and a scatter plot with legible axes in Figure 8. Please also address the denominator problem: reports per prompt explored is mechanically inflated when time spent reading examples displaces prompting inside a fixed 40-minute session. Reporting the raw counts alongside the ratio would help.
6. §6.1: drop the 82.6% and 74% pairing as support, or explain it. Per unit of time, the comparison page produced proportionally fewer reports than the rest of the interface, which is the opposite of what the sentence implies.
7. Split §6.2.2 by session. With $N_1=15$ and $N_2=30$ and a feature that displays what previous auditors tagged, the second cohort saw a populated board and the first did not. Report the 12/45 figure as two numbers.
8. Fix §3.1.1 against Table 3: six bachelor's degrees, not five; one high school, not two; one associate's, not two; and "Some college, no degree" (U06) is missing from the prose. Table 3 also codes U08 as "Man" where every other row uses "Female", and its caption reads "reported reported".
9. Table 5 caption versus §7: one startup or two.
10. Give the Stable Diffusion configuration in §4.2.8, including the images-per-prompt count, and say whether Replicate's safety filter was active.
11. Test the marginalization hypothesis in §6.3.2 instead of leaving it as a conjecture. The identity checkbox from §4.2.5 and the verifier demographics from §5.1.1 are enough for a cross-tabulation, and if the cell counts are too small to support one, say that explicitly.
12. Report what each group of participants was paid, in §5.1 and §5.2, and connect it to the argument in §8.1.
13. Release the 50 worked examples, the tag vocabularies, the verification survey, and the coding scheme as an appendix or a repository, and restore the system link now that the paper is not under blind review.

14. Reframe "three-week user study" in the abstract and §5 as a 90-minute session plus an optional three-week forum period, and move the 17-of-45 continuation rate into §8.1 as a result rather than leaving it in §6.4 as background.
15. Table 1 is referenced in §4.1 as "Table 1 above" but appears below the reference. Minor. While in there: §6.2.3 reports 40 of 167 as 23.9% where it rounds to 24.0%, §2.2 has "Fianlly", §3.1.1 has "system", §4.2.4 has "initial evidences", §4.2.8 writes "SQLi DB" where SQLite is presumably meant, §7 has "evaluate an improved version of WeAudit to with a broader population", and §8.2.1 has "possibility tailored".
16. One framing suggestion rather than a correction. §6.4.1's engaged commenters (nine frequent commenters, seven of whom filed only one or two reports) is the freshest empirical observation in the paper, and it is currently buried in a subsection. It bears directly on §8.1.1's argument about diverse roles. Consider promoting it.

5. Recommendation

Major revisions.

I want to be clear about why, because the verdict is not about the contribution. The design goals are grounded in real formative work, the system is a substantial build, the study collected a rich and varied dataset, and the practitioner interviews are the part of this literature that usually goes missing. The engaged-commenter observation and the invisible-labor discussion in §6.5 are both things I had not seen framed this way before, and the paper's own limitations section is candid where many are not.

What holds it back is that the claims in the abstract and in §1 are stated comparatively while the design supports only description, and that several reported quantities do not reconcile: the verifier demographics against their own total, the follow-up count against itself, the formative education breakdown against Table 3, the startup count against Table 5. The σ in §6.3.2 is the one I consider serious, since it makes a subsection contradict itself and that subsection carries an argument the discussion builds on. None of this requires new fieldwork. Recomputing the verification statistics, splitting two results by session, restating the comparative verbs, and publishing the materials would get this to acceptance. If the authors are willing to add the one no-scaffolding arm suggested in item 4, the actionability claim becomes something a reader can believe rather than something practitioners said in a demo, and the paper gets considerably stronger for it.

Declarations

Potential competing interests: No potential competing interests to declare.