Skepthical review: *The Reputational Tax of AI: How Structural Support and Incentives Shape Employee Disclosure Behavior*

Summary

The paper studies strategic misreporting of workplace AI use—either concealment (downplaying AI reliance) or performative disclosure (overstating AI use)—and argues that both can arise from a "reputational tax" created by algorithmic uncertainty (Introduction; Sec. 1; Sec. 2.1–2.3). Using survey data from **2,395** "Active AI Users," the authors model three disclosure outcomes (Transparent Reporting vs Concealment vs Performative Disclosure) via multinomial logit, relating these behaviors to perceived AI error frequency, job security confidence, agentic AI deployment (and an error × agentic interaction), structural "Foundational Support," cultural factors (learning safety; candid communication), and intrinsic/extrinsic incentives (Sec. 2.2–2.4; Sec. 3.3–3.6). The main findings are that perceived AI error frequency is associated with both concealment and performative disclosure (Sec. 3.3), concealment is additionally linked to job insecurity and is amplified under agentic AI deployment (Sec. 3.4), performative disclosure is strongly associated with intrinsic (but not extrinsic) rewards (Sec. 3.5), and Foundational Support robustly reduces both forms of misreporting while cultural measures are weaker/mixed (Sec. 3.6). The topic is timely and the concealment vs performative taxonomy is useful, but the paper's claims and policy implications would be substantially strengthened by (i) much fuller measurement transparency/construct validation—especially for self-admitted misreporting and composite indices—(ii) more careful handling of endogeneity/common-method bias and causal language, and (iii) additional robustness around the multinomial choice structure (IIA) and the central agentic-AI interaction (Sec. 2–4).

Strengths

- Timely and practically important question: strategic reporting of AI use can distort organizational AI ROI measurement and governance (Introduction; Sec. 4).
- Conceptually helpful distinction between defensive concealment and opportunistic performative disclosure; the taxonomy is clear and sustained through hypotheses and results (Introduction; Sec. 3.3–3.5).
- Large sample of active AI users and a generally appropriate initial modeling choice (multinomial logit) for a three-category outcome (Sec. 2.1; Sec. 2.3–2.4).
- Action-oriented framing that compares structural supports vs cultural initiatives, which is likely to resonate with practitioners (Sec. 3.6; Sec. 4).
- The analysis includes some standard diagnostics and robustness elements (VIF; IIA test; staged models; interaction testing) and the manuscript is internally numerically consistent in many places (Sec. 3.6.1).

- The results surface several nuanced patterns (error frequency relating to both misreporting directions; agentic moderation for concealment; intrinsic vs extrinsic incentives separating) that are plausible and theoretically generative (Sec. 3.3–3.5).

Major issues

1. **Outcome (misreporting) measurement validity and reproducibility are currently the largest threat to the paper's contribution. The dependent variable appears to rely on self-admitted concealment/exaggeration, but the exact item wording, response options, and classification rules into the three mutually exclusive categories (Transparent vs Concealment vs Performative) are under-specified (Sec. 2.2.1; Sec. 3.2–3.3).** This also creates a conceptual ambiguity: "Transparent Reporting" may effectively mean "does not admit misreporting," not necessarily accurate disclosure. Social desirability and introspective limits may bias both prevalence and estimated associations, and could affect the interpretation of coefficients as 'drivers' of misreporting rather than correlates of willingness to admit misreporting (Abstract; Introduction; Sec. 3.3–3.6).

   Recommendation: In Sec. 2.2.1, fully document the DV: exact question(s), response scale/anchors, whether items were single- or multi-item, and the deterministic rules/thresholds used to map responses into the three categories (including how ties/ambiguous responses were handled). Add an appendix table listing DV items and coding logic. In Sec. 4 (limitations), explicitly discuss (i) social desirability/common-method concerns, (ii) the possibility that the baseline category captures 'non-admission' rather than truth, and (iii) the likely direction(s) of bias. If feasible with existing data, add robustness checks aligned with the conceptual structure: e.g., a two-stage/nested formulation (misreport vs not; then conceal vs perform) or alternative codings (e.g., collapsing to misreport vs transparent; or treating intensity/ordinality if the response scale supports it), and report whether key conclusions hold.

2. **Key predictor measurement/coding is insufficiently detailed for replication and interpretability, especially for central theoretical constructs: Perceived AI Error Frequency, Job Security Confidence, Agentic AI Deployment, and Intrinsic/Extrinsic Rewards (Sec. 2.2–2.2.2; Sec. 3.3–3.5).** Several are described only as numeric/categorical/binary without item wording, scale ranges, anchors, reference categories, or examples—yet the paper's story depends on how respondents interpreted terms like "agentic AI," "intrinsic rewards," and "extrinsic rewards" (Sec. 1; Sec. 3.4–3.5).

   Recommendation: Expand Sec. 2.2–2.2.2 with a measurement/codebook-style description for every predictor: item wording, response options (including anchors), coding (including reference categories for categorical variables), and any transformations (standardization/centering). For complex constructs (agentic AI; intrinsic/extrinsic rewards), include brief definitions and concrete examples exactly as presented to respon-

dents; clarify whether items refer to team vs organization vs tool-level deployment. Report prevalence rates (percent "yes") for binary items in Sec. 3.1 or at the start of Sec. 3.4–3.5 to contextualize interactions and null findings.

3. **Causal/mechanistic language and policy claims are too strong for a cross-sectional, same-source survey with modest explained variance (Pseudo $R^2 \approx .078$–$.086$) and plausible endogeneity/common-method bias (Abstract; Introduction; Sec. 3.3–3.6; Sec. 4).** For example, claims that Foundational Support 'mitigates' misreporting or is 'more effective' than culture may be confounded by unobserved organizational maturity, leadership quality, compliance posture, AI governance, role mix, and individual traits (conscientiousness/anxiety) that influence both perceived support and reporting behavior.

   Recommendation: Across Abstract/Introduction/Results/Conclusion, systematically replace causal phrasing ("drives," "mitigates," "amplifies," "effective tool," "most powerful") with associational language ("is associated with," "predicts," "is consistent with"). Add a clearly labeled limitations subsection in Sec. 4 that foregrounds endogeneity, common-method bias, and omitted variables. If feasible with the dataset, strengthen identification/robustness by (i) adding richer controls (industry, job family, tenure, job level/manager status, AI tool type, AI maturity/governance proxies), (ii) clustering standard errors at the organization level (and/or including organization fixed effects if multiple respondents per firm and firm identifiers exist), and (iii) including a simple sensitivity analysis (e.g., Oster-style bounds/E-values) to quantify how strong unobserved confounding would need to be to eliminate the key Foundational Support association.

4. **Construction and conceptual separation of 'Foundational Support' (structural) vs cultural indices (Learning Safety; Candid Communication) is under-specified and may not cleanly support the "structure beats culture" conclusion (Sec. 2.2.2; Sec. 3.2; Sec. 3.6; Sec. 4).** PCA is used after CFA non-convergence, but the manuscript does not provide full item lists, loadings, cross-loadings, dimensionality checks, or reliability statistics for all indices. The first component explaining $\sim$ **31%** variance suggests potential multidimensionality; overlap between structural and cultural items could also attenuate or mask cultural effects (including the null for learning safety).

   Recommendation: In Sec. 2.2.2 and Sec. 3.2, report (preferably in an appendix) the full item lists, response scales, factor/PCA loadings (and cross-loadings if applicable), communalities, and Cronbach's α/ω for Foundational Support, Learning Safety, and Candid Communication. Justify PCA vs EFA vs averaged-scale construction given the theoretical intent (latent construct vs index). To substantiate the structure/culture distinction, run an EFA (or factor analysis) including all relevant items together and report whether distinct factors emerge; if they do not, qualify the structural-vs-cultural conclusion in Sec. 3.6 and Sec. 4 as contingent on these operationalizations.

5. **Model specification reporting is incomplete, limiting evaluation and replication (Sec. 2.3–2.4; Sec. 3.3–3.6.1). The staged models are described narratively without a compact specification table; reference categories for categorical predictors are not always explicit; and readers do not have full coefficient/SE output for the multinomial models. Standardization is described as "fully standardized," but the procedure is not defined for multinomial logit (Secs. 2.4 and 3.6; Figure 3).**

   Recommendation: Add (i) a model-specification table in Sec. 2.3–2.4 enumerating each stage's predictors, coding, and reference categories, and (ii) full regression tables (main text or appendix) for every multinomial model: coefficients, SEs, z, p, N, log-likelihoods, pseudo-R^2, and outcome baseline category (Transparent). Define "fully standardized" precisely (which variables were z-scored; how binary/categorical predictors were handled; whether centering was applied; any rescaling of coefficients) and provide the formula used for Figure 3's standardized coefficients.

6. **The multinomial logit IIA assumption and the reported Hausman–McFadden result ("HM Stat = 0.0, $p = 1.0$") are not adequately explained given the conceptual similarity of concealment and performative disclosure as related 'misreporting' choices (Sec. 2.4; Sec. 3.6.1).** The HM statistic can be numerically unstable/uninformative; without details (which alternative omitted, any warnings), it is hard to assess robustness.

   Recommendation: In Sec. 3.6.1 (or Sec. 2.4), explicitly document the IIA test procedure: which category was dropped, the exact test statistic format ($\chi^2(df)$), and whether any numerical issues occurred. Add at least one robustness check that relaxes IIA and/or matches the paper's conceptual structure: a nested logit (stage 1: misreport vs transparent; stage 2: conceal vs perform), multinomial probit, or a two-equation approach. If infeasible, acknowledge IIA as a limitation and argue substantively why modest violations are unlikely to change the headline findings.

7. **The 'agentic AI deployment' interaction is central to the narrative (Agentic Shift), but construct validity and interpretation need strengthening (Sec. 1; Sec. 3.4).** A binary indicator with unclear scope (team vs organization) and respondent knowledge may be misclassified; agentic deployment likely correlates with AI maturity, governance, and role types—plausible confounds for concealment and job insecurity.

   Recommendation: Provide the exact definition and item wording for agentic deployment (Sec. 2.2.2), including examples of "agentic" vs non-agentic tools and the organizational level referenced. In Sec. 3.4, report subgroup sizes/prevalence and present marginal effects or predicted probabilities (with confidence bands) for the interaction (error frequency × agentic) to make the substantive size interpretable. Add robustness

checks with additional controls (industry/job family/job level/AI maturity proxies) and/or organization-level clustering/FE (if available) to show the interaction is not an artifact of correlated organizational characteristics.

8. **The incentives result (intrinsic rewards associated with performative disclosure; extrinsic not significant) is intriguing but currently under-identified and risks over-interpretation as "innovation theater" (Sec. 3.5; Sec. 4).** Intrinsic rewards may proxy AI discourse salience or innovation-oriented teams where overclaiming is more visible/admitted; extrinsic rewards nulls may reflect low prevalence or coarse (binary) measurement; reverse causality is also plausible (people who overclaim may perceive/endorse intrinsic rewards).

   Recommendation: In Sec. 3.5, report prevalence for intrinsic/extrinsic reward measures and provide effect sizes as predicted probabilities (not just log-odds). Temper mechanism claims in Sec. 4 to "consistent with" rather than definitive theater/strategic gaming. If feasible, (i) test interactions (e.g., intrinsic rewards × job security; intrinsic × cultural climate) to assess boundary conditions; (ii) add controls for organizational AI maturity and role type; and (iii) discuss reverse-causality possibilities explicitly in limitations.

Minor issues

1. Sampling/recruitment and external validity are not described in sufficient detail, and selection into the "Active AI Users" subsample may limit generalizability (Sec. 2.1; Sec. 3.1; Sec. 4). The paper notes demographic differences between Active and Low-Frequency users but does not integrate these into interpretation or robustness checks; misreporting dynamics among infrequent/non-users could be particularly relevant to the performative-disclosure story.

   Recommendation: Augment Sec. 2.1 with sampling frame, recruitment channels/panel provider(s), geography/industry coverage, response rate (if available), and inclusion/exclusion criteria. In Sec. 3.1 and Sec. 4, explicitly narrow claims to "among employees using AI at least weekly" unless analyses support broader generalization. If feasible, run supplementary models on the full sample including usage frequency (and/or a selection model/weights) to assess whether relationships differ for low-frequency users.

2. Figures (esp. Figures 1–3) emphasize log-odds and standardized coefficients, which reduces interpretability for many readers; some captions/labels lack reference categories, scaling notes, and uncertainty visualization (notably Figure 2), and there are accessibility concerns due to reliance on color and small text (Sec. 3.3–3.6).

   Recommendation: Where possible, present effects as odds ratios and/or predicted probabilities/marginal effects alongside coefficient plots. Ensure every figure caption states (i) outcome baseline category (Transparent), (ii) predictor reference categories,

(iii) scaling/standardization, and (iv) CI construction. Add uncertainty (CIs/bands) to Figure 2, fix any rendering/label truncation issues, use colorblind-safe palettes with redundant encodings, and enlarge text for publication legibility.

3. Literature positioning could be stronger: the "reputational tax" framing and the concealment vs performative distinction would benefit from tighter grounding in impression management/self-presentation, gaming/metrics, psychological safety, and reporting bias/measurement distortion literatures (Introduction; Sec. 4).

   Recommendation: Add a concise related-work subsection in the Introduction that situates the contribution relative to impression management and organizational reporting, psychological safety and voice, technology use/adoption, and measurement bias. Clearly state what is novel (the two-direction taxonomy; linking AI uncertainty to disclosure strategies; structural vs cultural levers) and what is borrowed.

4. Reporting of predictive performance is limited to pseudo-R^2 and LR tests; given the applied relevance ("measurement distortion"), readers may want a sense of practical predictability and calibration (Sec. 3.3–3.6).

   Recommendation: Add brief predictive diagnostics: e.g., confusion matrix/classification accuracy (with caveats), Brier score/log-loss, or cross-validated out-of-sample performance. At minimum, contextualize pseudo-R^2 and emphasize what the model can/cannot predict in practice.

5. Ethical considerations are underdeveloped given the sensitivity of measuring misreporting and the risk of managerial misuse (Sec. 2–4).

   Recommendation: In Sec. 2 or Sec. 4, add a short ethics paragraph describing anonymity/confidentiality protections, whether IRB/ethics approval was obtained, and cautioning against punitive monitoring; emphasize that implications should focus on improving support structures and governance rather than sanctioning individuals.

6. Some quantitative claims are difficult to verify because key intermediate statistics are missing (e.g., null-model log-likelihood for the main LLR; incomplete coefficient/SE reporting in the Foundational Support validation regression) (Sec. 3.2–3.3).

   Recommendation: Report the intercept-only/null log-likelihood(s) needed to recompute LR tests, and provide SEs/t-stats for the Foundational Support validation regression coefficients. Ensure reported χ^2 degrees of freedom match the exact parameter count implied by the final model specification.

Very minor issues

1. Minor formatting/notation issues reduce polish: stray markdown markers, inconsistent p-value/Cronbach's α formatting, and mixed use of internal variable codes (e.g., QGS_Num, QGR) without a crosswalk (Sec. 1–3; figure captions).

Recommendation: Perform a careful copy-edit to standardize headings, statistical notation, and terminology. Add a short variable crosswalk table (survey item codes ↔ analytic variable names) in an appendix or footnote for quick reference.

2. Equation (1) presentation is slightly incomplete/ambiguous (index sets and implied probability normalization are not explicitly stated) (Sec. 2.3–2.4).

   Recommendation: Add a sentence specifying the set of outcome indices (e.g., $j \in$ **Concealment, Performative**) and include (or reference) the implied softmax probability expression so readers can map coefficients to predicted probabilities unambiguously.

3. Several cross-references are minimally descriptive (e.g., "visualized in Figure X"), and the IIA statistic label "HM Stat" is opaque (Sec. 3.3–3.6.1).

   Recommendation: When referencing figures, add a short descriptor of what the figure displays (e.g., "predicted probabilities by agentic deployment"). Replace "HM Stat" with "Hausman–McFadden test: $\chi^2(df) = \ldots$, $p = \ldots$" to match conventional reporting.

Key statements and references

- • **The multinomial logistic regression showed that perceived AI error frequency significantly increased the log-odds of both Concealment ($\beta = 0.160$, $p = .036$) and Performative disclosure ($\beta = 0.391$, $p < .001$), while higher levels of Foundational Support significantly reduced the log-odds of Concealment ($\beta = -0.310$, $p < .001$) and Performative disclosure ($\beta = -0.285$, $p < .001$), indicating that algorithmic uncertainty drives misreporting whereas structural support mitigates it.**
- *Reference(s):* (none)
- • **An interaction model including Agentic AI Deployment and perceived AI error frequency provided a significantly better fit than the main-effects model (Log-Likelihood $= -1394.9$, Pseudo $R^2 = .086$, LLR $\chi^2(2) = 23.4$, $p < .001$), with a positive, statistically significant interaction for Concealment ($\beta = 0.384$, $p = .028$) but not for Performative disclosure ($\beta = 0.013$, $p = .929$), showing that the reputational impact of AI errors on concealment is amplified specifically when agentic AI systems are deployed.**
- *Reference(s):* (none)
- • **In the sensitivity analysis of organizational incentives, Extrinsic Rewards for AI usage had no significant effect on Performative disclosure ($\beta = 0.026$, $p = .856$), whereas the availability of Intrinsic Rewards was a strong positive predictor ($\beta = 0.996$, $p < .001$), and a likelihood ratio test confirmed that including Intrinsic Rewards yielded a far superior model fit compared**

to a model with only Extrinsic Rewards ($\chi^2(2) = 48.62$, $p < .001$), indicating that social, non-monetary incentives strongly promote performative overstatement of AI use.

- *Reference(s):* (none)

Mathematical consistency audit

This section audits **symbolic/analytic** mathematical consistency (algebra, derivations, dimensional/unit checks, definition consistency).

Maths relevance: light

The paper contains limited formal mathematics: one multinomial logistic regression log-odds equation and several references to standard statistical constructs (PCA indices, Cronbach's α, LR tests, pseudo-R^2, VIF, Hausman–McFadden IIA test). The main analytic consistency questions are about parameter counting/df reporting for LR tests and the undefined procedure for "fully standardized" multinomial logit coefficients.

Checked items

1. ✔ **Multinomial log-odds model form** (Eq. (1), Sec. 2.3, p.4)
    - **Claim:** For each outcome j relative to Transparent reporting, $\log\left(P(Y_i = j)/P(Y_i = \text{Transparent})\right)$ is linear in predictors with outcome-specific coefficients.
    - **Checks:** algebra/definition consistency, notation consistency
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** Transparent Reporting is the reference outcome, Outcomes are unordered and modeled via multinomial logit, Coefficients β_j are outcome-specific for each non-reference category
    - **Notes:** Eq. (1) is the standard baseline-category multinomial logit log-odds specification. Notation is consistent with the stated reference category.
2. ✔ **Dependent variable category coding vs reference outcome** (Sec. 2.2.1, p.3 and Sec. 2.3, p.4)
    - **Claim:** Transparent Reporting is coded as **0** and is used as the baseline/reference category in the multinomial logit.
    - **Checks:** definition consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Coding **0/1/2** is arbitrary but baseline must match reference in the model
    - **Notes:** The baseline category is clearly stated as Transparent Reporting (**0**) and is used as the reference in Eq. (1).

3. ✔ **Index construction via PCA and stated distribution** (Secs. 2.2.2 and 3.2, pp.3 and 6)
    - **Claim:** Foundational Support is the first principal component and is approximately normally distributed with mean near 0; used as a predictor in the models.
    - **Checks:** definition consistency, sanity/limiting checks
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** PCA scores can be centered at 0 depending on scoring/standardization, Using PC1 as an index is a linear combination of items
    - **Notes:** Nothing mathematically contradictory: a PC score commonly has mean ~ 0 if computed from centered variables. The SD value depends on scaling choices; the paper does not specify scaling but the claim is not self-contradictory.
4. ✔ **Regression of Foundational Support on size and revenue: df consistency** (Sec. 3.2, p.6)
    - **Claim:** Regressing the index on two predictors yields $F(2, 2392)$, consistent with the sample size.
    - **Checks:** degrees-of-freedom consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Analytical sample size is $N = 2395$, Model includes 2 predictors plus intercept
    - **Notes:** For an OLS regression with $N = 2395$ and $k = 2$ predictors, denominator df is $N - k - 1 = 2392$, matching the reported $F(2, 2392)$.
5. ⚠ **Main multinomial model LR test df plausibility** (Sec. 3.3, p.6)
    - **Claim:** Overall model LLR $\chi^2(10)$ corresponds to the fitted multinomial model described in Sec. 3.3.
    - **Checks:** degrees-of-freedom consistency, model specification completeness
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Multinomial logit with 3 categories has 2 non-reference equations, LR-test df should equal the number of estimated slope parameters added relative to the null (intercepts-only) model
    - **Notes:** If the model truly contains only 5 predictors total (e.g., error frequency + foundational support + 3 job-security dummies), then $df = 10$ is consistent (2 equations $\times$ 5 slopes). However, the paper's narrative and figure labels suggest Job Security Confidence may have more than 4 total categories (implying ≥ 4 dummies) or additional predictors, which would change df. The exact dummy coding and full predictor list are not shown, preventing verification.

6. ⚠ **Interaction term interpretation vs model form** (Sec. 3.4, pp.7–8)
    - **Claim:** An interaction between Agentic AI Deployment (binary) and Perceived AI Error Frequency modifies the effect of error frequency on concealment/performative outcomes.
    - **Checks:** definition consistency, model specification completeness
    - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
    - **Assumptions/inputs:** Interaction included as product term (Agentic × Error) in the linear predictor for each non-reference outcome, Main effects of Agentic and Error are also included (not explicitly stated but typically required for interpretability)
    - **Notes:** The paper states an interaction term was introduced and reports its coefficient for each outcome, which is consistent at a high level. But the exact equation (including whether both main effects are retained) is not written, so the precise specification cannot be audited.
7. ⚠ **"Fully standardized" multinomial model coefficients** (Secs. 2.4 and 3.6, pp.5 and 9; Figure 3 caption pp.10)
    - **Claim:** A fully standardized model allows comparing effect sizes on a common scale; standardized coefficients for Foundational Support and Learning Safety are reported.
    - **Checks:** definition completeness, notation/interpretation consistency
    - **Verdict:** UNCERTAIN; confidence: high; impact: moderate
    - **Assumptions/inputs:** Standardization method must be defined to interpret magnitudes, For categorical/binary predictors, standardization choices affect scale
    - **Notes:** The term "fully standardized" is not defined, and multinomial logit does not have a single canonical standardization. Without an explicit procedure, the reported standardized β values cannot be mathematically verified or compared unambiguously.

Limitations

- Only one explicit equation is provided; most claims are verbal descriptions of statistical models and results, limiting the scope of symbolic verification.
- The audit cannot confirm parameter counting for LR tests or dummy-variable construction because the full model design matrix (or coefficient table) is not included in the provided PDF text.
- Figures contain variable codes (e.g., QGS_Num, QGR) but the paper does not provide a complete symbol/variable mapping, limiting notation consistency checks.

Numerical results audit

This section audits **numerical/empirical** consistency: reported metrics, experimental design, baseline comparisons, statistical evidence, leakage risks, and reproducibility.

Of **22** candidate numeric checks, **19** passed and **3** were uncertain (none failed). Recomputations confirmed multiple reported p-values (t-tests and χ^2 tests), the F-test p-value for the validation regression, and the likelihood-ratio statistic derived from paired log-likelihoods for the interaction model. Remaining uncertainties are driven by missing inputs needed for recomputation (e.g., standard errors and null-model log-likelihood).

Checked items

1. ✔ **C1_sample_counts_total** (p.3 (Methods 2.1) and p.5 (Results 3.1))
    - **Claim:** Filtering yielded a final sample of **2,395** Active AI Users; a comparative group of **208** Low-Frequency Users remained.
    - **Checks:** parts_vs_total (compute implied total N)
    - **Verdict:** PASS
    - **Notes:** Implied total respondents $= 2395 + 208 = 2603$ (active share ≈ 0.9201; low-frequency share ≈ 0.0799). No reported overall total was provided to compare against.
2. ✔ **C2_pca_variance_range** (p.3 (Methods 2.2.2) and p.6 (Results 3.2))
    - **Claim:** The first principal component explained **31.12%** of the variance.
    - **Checks:** range_check (percentage bounds)
    - **Verdict:** PASS
    - **Notes:** 31.12% is within $[0, 100]$.
3. ✔ **C3_cronbach_alpha_range** (p.3 (Methods 2.2.2) and p.6 (Results 3.2))
    - **Claim:** Foundational Support index internal consistency: Cronbach's $\alpha = .797$.
    - **Checks:** range_check (Cronbach alpha bounds)
    - **Verdict:** PASS
    - **Notes:** 0.797 is within $[0, 1]$.
4. ✔ **C4_index_mean_sd_sanity** (p.6 (Results 3.2))
    - **Claim:** Foundational Support index approximately normally distributed ($M = 0.0$, $SD = 1.933$).
    - **Checks:** range_check (SD nonnegative)
    - **Verdict:** PASS
    - **Notes:** $SD = 1.933$ is nonnegative; the normality claim itself is not verifiable from summary stats alone.
5. ✔ **C5_regression_df_matches_n** (p.6 (Results 3.2))

- **Claim:** Regression validating Foundational Support: $F(2, 2392) = 3.00$ with sample size 2,395.
- **Checks:** degrees_of_freedom_consistency
- **Verdict:** PASS
- **Notes:** df_{resid} matches OLS with intercept: $2392 = 2395 - 2 - 1$.

6. ✔ **C6_r2_in_0_1** (p.6 (Results 3.2))
   - **Claim:** Model explained negligible variance ($R^2 = .0025$).
   - **Checks:** range_check (R^2 bounds)
   - **Verdict:** PASS
   - **Notes:** $R^2 = 0.0025$ is within $[0, 1]$.
7. ✔ **C7_f_p_value_consistency** (p.6 (Results 3.2))
   - **Claim:** Reported $F(2, 2392) = 3.00$, $p = .050$.
   - **Checks:** p_value_recompute_from_F
   - **Verdict:** PASS
   - **Notes:** Recomputed $p = 0.0499744333$, consistent with reported $p = 0.050$ within tolerance.
8. ⚠ **C8_beta_p_consistency_employee_size** (p.6 (Results 3.2))
   - **Claim:** Employee size predictor: $\beta = 0.057$, $p = .075$ (non-significant).
   - **Checks:** p_value_recompute_from_beta_as_t
   - **Verdict:** UNCERTAIN
   - **Notes:** Cannot recompute p-value from β alone without standard error (or t -statistic).
9. ⚠ **C9_beta_p_consistency_annual_revenue** (p.6 (Results 3.2))
   - **Claim:** Annual revenue predictor: $\beta = 0.004$, $p = .921$.
   - **Checks:** p_value_recompute_from_beta_as_t
   - **Verdict:** UNCERTAIN
   - **Notes:** Cannot recompute p-value from β alone without standard error (or t -statistic).
10. ✔ **C10_age_ttest_p_value** (p.5 (Results 3.1))
    - **Claim:** Age difference: Active users $M = 36.19$ $SD = 9.28$ vs Low-Frequency $M = 42.12$ $SD = 12.73$, $t(226.51) = -6.56$, $p < .001$.
    - **Checks:** p_value_recompute_from_t
    - **Verdict:** PASS
    - **Notes:** Two-sided p recomputed from t and df is $3.608860888945861 \times 10^{-10}$, satisfying $p < .001$.
11. ✔ **C11_experience_ttest_p_value** (p.5 (Results 3.1))

- **Claim:** Years in workforce: Active $M = 10.96$ $SD = 8.79$ vs Low-Frequency $M = 18.41$ $SD = 12.89$, $t(224.04) = -8.17$, $p < .001$.
- **Checks:** p_value_recompute_from_t
- **Verdict:** PASS
- **Notes:** Two-sided p recomputed from t and df is $2.2548303799537694 \times 10^{-14}$, satisfying $p < .001$.

12. ✔ **C12_chi_square_p_value_roles** (p.5 (Results 3.1))
    - **Claim:** Occupational roles difference: $\chi^2(10) = 71.07$, $p < .001$.
    - **Checks:** p_value_recompute_from_chi_square
    - **Verdict:** PASS
    - **Notes:** Recomputed $p = 2.7541759878139427 \times 10^{-11}$, satisfying $p < .001$.
13. ✔ **C13_chi_square_p_value_industry** (p.5 (Results 3.1))
    - **Claim:** Industry sectors difference: $\chi^2(12) = 90.45$, $p < .001$.
    - **Checks:** p_value_recompute_from_chi_square
    - **Verdict:** PASS
    - **Notes:** Recomputed $p = 4.039755101542551 \times 10^{-14}$, satisfying $p < .001$.
14. ✔ **C14_chi_square_p_value_employee_size_nonsig** (p.5 (Results 3.1))
    - **Claim:** No difference in Global Employee Size: $\chi^2(5) = 4.31$, $p = .505$.
    - **Checks:** p_value_recompute_from_chi_square
    - **Verdict:** PASS
    - **Notes:** Recomputed $p = 0.5056997255$, consistent with reported $p = 0.505$ within tolerance.
15. ✔ **C15_chi_square_p_value_revenue_nonsig** (p.5 (Results 3.1))
    - **Claim:** No difference in Global Annual Revenue: $\chi^2(4) = 6.48$, $p = .166$.
    - **Checks:** p_value_recompute_from_chi_square
    - **Verdict:** PASS
    - **Notes:** Recomputed $p = 0.1660549152$, consistent with reported $p = 0.166$ within tolerance.
16. ⚠ **C16_llr_matches_loglik_diff_main_model** (p.6 (Results 3.3))
    - **Claim:** Overall model fit: Log-Likelihood $= -1406.6$; LLR $\chi^2(10) = 238.6$.
    - **Checks:** likelihood_ratio_from_loglik_difference (requires null LL)
    - **Verdict:** UNCERTAIN
    - **Notes:** Cannot compute LLR without the null-model log-likelihood.
17. ✔ **C17_agentic_interaction_llr_from_loglik_diff** (p.7 (Results 3.4) and p.6 (Results 3.3))

- **Claim:** Interaction model better fit: Log-Likelihood = -1394.9 vs main effects model Log-Likelihood = -1406.6; LLR $\chi^2(2) = 23.4$.
- **Checks:** likelihood_ratio_from_two_logliks
- **Verdict:** PASS
- **Notes:** Computed LLR $= 2 * ((-1394.9) - (-1406.6)) = 23.4$, matching the reported LLR to rounding.

18. ✔ **C18_agentic_interaction_p_value_from_llr** (p.7 (Results 3.4))
    - **Claim:** Interaction model LLR $\chi^2(2) = 23.4$, $p < .001$.
    - **Checks:** p_value_recompute_from_chi_square
    - **Verdict:** PASS
    - **Notes:** Recomputed $p = 8.293819160757377 \times 10^{-6}$, satisfying $p < .001$.
19. ✔ **C19_intrinsic_vs_extrinsic_lrtest_p_value** (p.8-9 (Results 3.5))
    - **Claim:** Likelihood ratio test: $\chi^2(2) = 48.62$, $p < .001$.
    - **Checks:** p_value_recompute_from_chi_square
    - **Verdict:** PASS
    - **Notes:** Recomputed $p = 2.768860940838673 \times 10^{-11}$, satisfying $p < .001$.
20. ✔ **C20_cultural_vars_lrtest_p_value** (p.9 (Results 3.6))
    - **Claim:** Adding cultural variables improved fit: Likelihood Ratio $\chi^2(4) = 27.48$, $p < .001$.
    - **Checks:** p_value_recompute_from_chi_square
    - **Verdict:** PASS
    - **Notes:** Recomputed $p = 1.5896126206410667 \times 10^{-5}$, satisfying $p < .001$.
21. ✔ **C21_hausman_mcfadden_p_value_from_stat** (p.10 (Figure 3 caption / robustness text))
    - **Claim:** Hausman-McFadden test: HM Stat = 0.0, $p = 1.0$.
    - **Checks:** p_value_consistency_sanity
    - **Verdict:** PASS
    - **Notes:** Sanity checks passed: HM stat is nonnegative; p is within $[0, 1]$; and stat $= 0$ is consistent with $p = 1$ under common chi-square-based constructions (df not reported, so only sanity check possible).
22. ✔ **C22_vif_threshold_claim** (p.9 (Results 3.6.1))
    - **Claim:** Maximum VIF = 1.54 and all VIFs below threshold of 5.
    - **Checks:** threshold_check
    - **Verdict:** PASS
    - **Notes:** 1.54 is below the stated threshold 5.

Limitations

- Only parsed text provided; no tables with full coefficient/SE outputs, no raw data, and no null-model log-likelihoods, limiting recomputation of several fit statistics (LLR for main model, pseudo R^2).
- Figure numeric values (confidence intervals, marginal probabilities) are not available as explicit numbers in the text, and plot-pixel extraction is disallowed.
- Some inferential statistics (e.g., p-values tied to regression coefficients) require standard errors or test statistics not reported.