Peer-Review

Gardos, Rafael-Petrut. 2026. “Conditional Uncertainty-Aware Political Deepfake Detection with Stochastic Convolutional Neural Networks.” *Journal of High School Science* 10 (1): 186–220. https://doi.org/10.64336/001c.156299.

1.It is unclear whether uncertainty improvements stem from Bayesian inference or simply from stochastic regularization. The improved calibration could simply be due to noise-induced smoothing rather than principled Bayesian uncertainty estimation. Please address this in the manuscritp by [1] Comparing MC dropout vs. single stochastic forward pass at test time. [2] Comparing against deep ensembles or temperature scaling, and measure uncertainty–error correlation (e.g., AUROC of uncertainty for error detection). [3] Perform ablation studies isolating noise vs. epistemic uncertainty.
2.Clearly frame the term “interpretability” as 'decision-level interpretability". This is because uncertainty estimation ≠ interpretability and you are over-reaching conceptually.
3.Accuracy vs. Calibration Trade-off Not Quantified. You state that trustworthiness improves “without sacrificing accuracy,” but:no quantitative trade-off is reported. It is unclear whether accuracy differences are statistically significant or within noise margins. In high-stakes applications, small accuracy losses may still matter.
4.Unclear Dataset Construction and Generalization. While the dataset is described as “balanced” and derived via a “custom filtering pipeline,” you do not clarify the number of samples, whether political identities, manipulation methods, or generation models overlap between train and test splits, whether results reflect in-distribution or cross-manipulation generalization. Coverage across politicians, events, imaging conditions, and generative models is not quantified; range limitations may explain low AUC. This raises concerns about potential dataset leakage or overly optimistic uncertainty estimates.
5.There are serious internal inconsistencies: e.g., EfficientNet-B4 ROC–AUC ~0.396 contradicts the reported confusion matrix that implies near-perfect accuracy; some confusion matrix counts exceed the stated test set size. This is a critical validity concern. At minimum, the Results section must explain: How thresholds were chosen. Whether ROC and confusion matrices are computed on the same split. Whether class labels or positive-class definitions differ. These results are statistically incompatible unless clarified. Possible explanations include: Class imbalance at evaluation time (contradicting earlier “balanced” claims). ROC computed on probabilities while confusion matrices reflect a fixed threshold with skewed outputs. Label inversion or reporting error. Multiple contradictions exist (AUC vs confusion matrix, totals vs test set size, calibration claims vs low AUC), undermining internal coherence. Check the entire manuscript.
6.Key assumptions (head-only training suffices for comparability; 380×380 for ResNet-18; avoiding ImageNet normalization) are not empirically justified and may harm transfer performance.
7.Images were scaled to [0,1] and ImageNet mean–std normalization was avoided; discuss implications for transfer learning and provide ablations with/without standard normalization.
8.Design compares deterministic vs Bayesian heads on two backbones with a balanced split, but lacks identity/generator stratification, OOD test sets, significance tests, and full reproducibility details (software versions, hardware, seeds, learning rates, batch sizes). There are no confidence intervals, hypothesis tests (e.g., DeLong for AUC), or uncertainty on metrics; paired comparisons and calibration curve uncertainties are absent. Potential confounders (identity overlap between splits, generator/source distribution, compression, and content context) are not controlled or statistically adjusted.
9.No sensitivity analyses (e.g., to T in MC dropout, dropout rate, input resolution), no cross-generator or compression robustness tests, and no perturbation/shift experiments are reported. Some conclusions are plausible (calibration vs discrimination are distinct), but they rest on contradictory results; inference about deployment safety is premature.
10.Misuse and Overconfidence in ROC-AUC Interpretation. You repeatedly describe AUC ≈ 0.39–0.55 as: “unexpectedly low”, “near chance”, “indicating weak global separability”. While broadly

correct, two scientific issues arise: AUC < 0.5 usually implies systematic misranking, not just weak discrimination. The discussion does not consider whether: Labels are inverted, Score distributions are skewed, Bayesian averaging alters score monotonicity. Without this analysis, conclusions about "weak separability" are incomplete.

11.Overinterpretation of Loss Curves: You infer from learning curves that: "EfficientNet-B4's pretrained features already encode rich semantic cues applicable to political imagery." This is not directly supported by loss trajectories alone. Loss convergence under frozen-backbone training shows:Optimization stability, Head-level separability. It does not demonstrate semantic alignment or task understanding.

12.Strong Language in Results Section Is Inappropriate. Results sections should report, not argue. Phrases such as:"This confirms that…""This demonstrates that…""The model understands…""Rendering it unusable", belong in the Discussion, not the Results. Their presence weakens scientific neutrality.

I deeply thank the reviewer for the detailed and technically rigorous feedback. Below, I reproduce each comment verbatim and describe precisely how and where it has been addressed in the revised manuscript.

1. "It is unclear whether uncertainty improvements stem from Bayesian inference or simply from stochastic regularization. The improved calibration could simply be due to noise-induced smoothing rather than principled Bayesian uncertainty estimation. Please address this in the manuscript by [1] Comparing MC dropout vs. single stochastic forward pass at test time. [2] Comparing against deep ensembles or temperature scaling, and measure uncertainty–error correlation (e.g., AUROC of uncertainty for error detection). [3] Perform ablation studies isolating noise vs. epistemic uncertainty."

I addressed this concern by explicitly reframing MC dropout as an empirically evaluated stochastic inference procedure rather than a guaranteed Bayesian epistemic estimator, and by introducing controls that separate "noise-induced smoothing" from multi-sample aggregation. Methods 3.F defines deterministic inference, single-pass stochastic inference (T=1), MC dropout (T>1) under identical training and evaluation semantics, and Methods 3.I defines uncertainty-error correlation via AUROC for error detection. In Results, the required comparisons are reported directly: single-pass stochastic inference (T=1) is reported alongside MC dropout (T=20 mean) and deterministic inference in 4.C and explicitly discussed as a control in 4.G; temperature scaling and the ensemble surrogate are included as non-dropout reference baselines in 4.C and further evaluated under generator-disjoint OOD conditions in 4.P; uncertainty-error correlation (AUROC for error detection) is quantified with entropy-and variance-based uncertainty scores in 4.I (tables and curves), and is contextualized as an operational signal rather than a Bayesian guarantee. Finally, ablations that isolate the effect of stochasticity intensity and Monte Carlo averaging are provided via MC sample-size sensitivity in 4.J and dropout-rate ablation in 4.K. It shows empirically that reliability behavior is not assumed to improve monotonically with increased sampling or increased dropout. For the interpretive implications (noise vs. aggregation vs. calibration baselines), I consolidated them in Discussion 5.B.

2. "Clearly frame the term "interpretability" as 'decision-level interpretability". This is because uncertainty estimation ≠ interpretability and you are over-reaching conceptually."

The manuscript now narrows the scope of "interpretability" explicitly to decision-level interpretability and removes any broader implication that uncertainty estimation provides mechanistic, feature-level or causal explanations. I stated that unambiguously in Methods 3.I ("Interpretability scope"), where interpretability is defined solely as the capacity of calibrated probabilities and uncertainty scores to support downstream decision-making (eg triage, abstention, escalation), and where non-decision interpretability claims are explicitly disclaimed. I kept the Results section descriptive and did not present uncertainty as an explanatory mechanism (4.A-4.Q), while the Discussion returns to interpretability only in this restricted operational sense (Discussion 5.C). It's aligning the conceptual framing with the empirical evaluation design.

3. “Accuracy vs. Calibration Trade-off Not Quantified. You state that trustworthiness improves “without sacrificing accuracy,” but:no quantitative trade-off is reported. It is unclear whether accuracy differences are statistically significant or within noise margins. In high-stakes applications, small accuracy losses may still matter.”

I quantified the accuracy-calibration relationship explicitly and attached uncertainty to the reported metrics to distinguish meaningful shifts from resampling noise. Methods 3.H defines the calibration-sensitive metrics (ECE, Brier, NLL) alongside accuracy under a fixed threshold, and Methods 3.E specifies bootstrap resampling and fixed evaluation semantics. Results 4.O reports bootstrap confidence intervals for accuracy and calibration metrics across inference procedures (deterministic, temperature scaling, T=1, MC T=20 mean, ensemble surrogate), enabling direct assessment of whether accuracy changes fall within sampling variability while calibration-sensitive metrics shift. Now, the Results wording is descriptive (reporting intervals rather than asserting “no sacrifice”), and Discussion 5.D interprets the trade-off in a high-stakes context by emphasizing that “small” accuracy deltas may still be operationally relevant and therefore must be reported with uncertainty rather than asserted categorically.

4. “Unclear Dataset Construction and Generalization. While the dataset is described as “balanced” and derived via a “custom filtering pipeline,” you do not clarify the number of samples, whether political identities, manipulation methods, or generation models overlap between train and test splits, whether results reflect in-distribution or cross-manipulation generalization. Coverage across politicians, events, imaging conditions, and generative models is not quantified; range limitations may explain low AUC. This raises concerns about potential dataset leakage or overly optimistic uncertainty estimates.”

I now made the dataset construction and generalization scope explicit, including sample counts, split sizes, the absence of identity/generator stratification in the primary split, and the resulting interpretation as matched-distribution (in-distribution) generalization rather than cross-identity or cross-generator generalization. Methods 3.A reports N=4000 total images with a balanced 2000/2000 class composition, and specifies the fixed image-level train/validation/test partitioning (2800/600/600). And something crucial, Methods 3.A states that the primary split is random and not stratified by political identity, event type, or generator, and adds an explicit statement that generator labels may appear in multiple splits such that the ID test evaluates matched-generator generalization rather than cross-generator generalization. To address the generalization concern directly rather than implicitly, Results 4.M reports generator-stratified evaluation within the ID test split (heterogeneity across generators that are present at test time), and Results 4.P reports a separate generator-disjoint OOD evaluation where unseen generators are held out from training. The limitations of identity-disjoint generalization and broader real-world shifts (platform compression pipelines, capture conditions, temporal drift) are explicitly bounded as out of scope for the main ID claims and are discussed as limitations in Discussion 5.E to prevent overgeneralization and leakage accusations.

5. “There are serious internal inconsistencies: e.g., EfficientNet-B4 ROC–AUC ~0.396 contradicts the reported confusion matrix that implies near-perfect accuracy; some confusion matrix counts exceed the stated test set size. This is a critical validity concern. At minimum, the Results section must explain: How thresholds were chosen. Whether ROC and confusion matrices are computed on the same split. Whether class labels or positive-class definitions differ. These results are statistically incompatible unless clarified. Possible explanations include: Class imbalance at evaluation time (contradicting earlier “balanced” claims). ROC computed on probabilities while confusion matrices reflect a fixed threshold with skewed outputs. Label inversion or reporting error. Multiple contradictions exist (AUC vs confusion matrix, totals vs test set size, calibration claims vs low AUC), undermining internal coherence. Check the entire manuscript.”

I resolved this validity concern by making evaluation semantics explicit (positive class definition, score direction, threshold rule) and by ensuring that ROC/AUC and confusion-matrix reporting are computed on the same split under consistent label conventions, eliminating the previously contradictory metric pairing. Methods 3.G now fixes the positive class as fake (y=1) throughout, defines the ROC score as $s(x)=p(y=1|x)$, and defines fixed-threshold decisions at $t=0.5$; it also states

that ROC curves and confusion matrices are computed on the same held-out test split unless otherwise stated. Results 4.A reports deterministic confusion matrices on the test split (n=600) and explicitly verifies that each confusion matrix sums to 600, preventing count-size inconsistency. Results 4.B reports ROC curves computed from the same score definition and split, and Results 4.C reports ROC-AUC numerically alongside the confusion-derived counts, so that ranking and threshold-based metrics are aligned and interpretable together. In addition, the Results text explicitly notes that AUC<0.5 would indicate systematic misranking (e.g. label inversion or score direction mismatch) and clarifies that this diagnostic does not apply under the corrected semantics (4.B), closing the loop on the specific misinterpretation risk raised by the reviewer. Discussion 5.A-5.B then interprets discrimination vs. reliability under these consistent semantics without relying on contradictory metric combinations.

6. "Key assumptions (head-only training suffices for comparability; 380×380 for ResNet-18; avoiding ImageNet normalization) are not empirically justified and may harm transfer performance."

I retrained the CNNs; no more head-only. I removed the head-only comparability assumption by stating that all reported results correspond to full end-to-end fine-tuning, and it treats preprocessing choices (resolution and normalization) as experimentally assessed factors rather than fixed assumptions. Methods 3.B explicitly states that the backbones are ImageNet-pretrained and are fully fine-tuned end-to-end for all reported results, avoiding frozen-backbone/head-only comparability claims. Methods 3.C acknowledges that using 380×380 for ResNet-18 and omitting ImageNet normalization deviates from canonical transfer learning practice and may affect transfer alignment; it therefore frames resolution and normalization as potential confounders that must be empirically tested rather than assumed harmless. The empirical justification is then provided in Results 4.L, which reports resolution and normalization ablations (including ImageNet mean-std normalization) with confidence intervals. It now allows you to see the direction and magnitude of any performance and calibration effects induced by these design choices. The implications and limits of these choices are discussed in Discussion 5.E, emphasizing that claims are conditional on the evaluated preprocessing configuration and that stronger generalization claims require broader shift testing.

7. "Images were scaled to [0,1] and ImageNet mean–std normalization was avoided; discuss implications for transfer learning and provide ablations with/without standard normalization."

I addressed this by (i) explicitly discussing the transfer-learning implications of deviating from ImageNet normalization, and (ii) adding explicit ablations that compare [0,1] scaling against ImageNet mean-std normalization (and dataset-specific normalization) under controlled evaluation semantics. Methods 3.C now states that omitting ImageNet normalization alters the input distribution relative to pretrained feature statistics and can affect both discrimination and calibration, and it points to a targeted empirical evaluation rather than implying neutrality. Results 4.L provides the required ablations across normalization strategies (including ImageNet mean-std) and resolutions with bootstrap confidence intervals, enabling direct comparison of discrimination (AUC), accuracy, and reliability metrics under the same split and label semantics. The Results wording is kept descriptive and does not assert universal superiority; it reports observed differences under the dataset distribution used, while in the Discussion 5.E, I clarify the conditional nature of these findings and their relevance for transfer alignment in politically filtered imagery.

8. "Design compares deterministic vs Bayesian heads on two backbones with a balanced split, but lacks identity/generator stratification, OOD test sets, significance tests, and full reproducibility details (software versions, hardware, seeds, learning rates, batch sizes). There are no confidence intervals, hypothesis tests (e.g., DeLong for AUC), or uncertainty on metrics; paired comparisons and calibration curve uncertainties are absent. Potential confounders (identity overlap between splits, generator/source distribution, compression, and content context) are not controlled or statistically adjusted."

I addressed the evaluation-design and reporting gaps by adding (a) explicit reproducibility details, (b) uncertainty quantification via bootstrap confidence intervals, (c) paired AUC significance testing via DeLong on a paired OOD evaluation, (d) generator-stratified reporting, and (e) generator-

disjoint OOD evaluation, while explicitly delimiting what remains uncontrolled (e.g., identity overlap) to prevent overclaiming. In Methods 3.E, I provide full reproducibility details, including software stack, hardware, training schedule, batch size, random seed control, and the evaluation protocol; Methods 3.A clarifies that the primary split is not identity- or generator-disjoint and therefore constitutes in-distribution matched-generator evaluation, and it defines the generator-disjoint OOD split used for robustness assessment. In Results, I reported metric uncertainty via bootstrap confidence intervals in 4.O, generator heterogeneity is reported via generator-stratified evaluation in 4.M, compression sensitivity is evaluated in 4.N, and generator-disjoint OOD performance is evaluated in 4.P. For statistical testing of discrimination differences under paired predictions, fortunately, Results 4.Q reports DeLong paired tests on OOD ROC-AUC, and the Results text notes the specific case in which DeLong is not applicable (temperature scaling preserves AUC by construction). Remaining confounders that are not statistically adjusted (identity overlap, correlated prompts/content) are explicitly bounded in Methods 3.A and discussed as limitations in Discussion 5.E rather than left implicit.

9. "No sensitivity analyses (e.g., to T in MC dropout, dropout rate, input resolution), no cross-generator or compression robustness tests, and no perturbation/shift experiments are reported. Some conclusions are plausible (calibration vs discrimination are distinct), but they rest on contradictory results; inference about deployment safety is premature."

In the Results section, I added targeted sensitivity and robustness analyses that explicitly test MC sample size (T), dropout rate, preprocessing sensitivity (resolution/normalization), generator heterogeneity, compression robustness, and generator-disjoint distribution shift, and it bounds deployment-facing claims accordingly. Sensitivity to MC sample size is reported in Results 4.J ($T \in \{1,5,10,20,50\}$) and sensitivity to dropout probability is reported in 4.K, both under fixed evaluation semantics. Preprocessing sensitivity to resolution and normalization is reported in 4.L with confidence intervals. Cross-generator heterogeneity within the matched-distribution test split is reported in 4.M, compression robustness is reported in 4.N, and distribution shift is operationalized via generator-disjoint OOD evaluation in 4.P with paired discrimination significance testing in 4.Q. In addition, I addressed the internal semantic consistency issues that previously undermined interpretability, as described in the response to Comment 5, so the new robustness/sensitivity conclusions are not built on contradictory metrics. Finally, Discussion 5.F explicitly constrains any deployment implications to a decision-support framing and avoids implying guaranteed safety under unconstrained real-world shifts.

10. "Misuse and Overconfidence in ROC-AUC Interpretation. You repeatedly describe AUC ≈ 0.39–0.55 as: "unexpectedly low", "near chance", "indicating weak global separability". While broadly correct, two scientific issues arise: AUC < 0.5 usually implies systematic misranking, not just weak discrimination. The discussion does not consider whether: Labels are inverted, Score distributions are skewed, Bayesian averaging alters score monotonicity. Without this analysis, conclusions about "weak separability" are incomplete."

I addressed this by explicitly incorporating the AUC<0.5 diagnostic interpretation into the Results narrative and by eliminating the prior mismatch between label/score conventions that could generate apparent AUC anomalies. Methods 3.G fixes the positive class, score definition, and threshold rule to prevent label inversion ambiguity, and Results 4.B explicitly states that AUC<0.5 indicates systematic misranking (e.g., inverted labels or reversed score direction) rather than merely weak discrimination, thereby aligning the interpretation with standard ROC theory. Because ROC/AUC is now computed consistently on $s(x)=p(y=1|x)$ and confusion matrices are computed under the same label semantics on the same split (4.A-4.C), the manuscript no longer relies on the earlier "near chance" narrative driven by inconsistent reporting; instead, it reports ROC curves and AUC values under the corrected semantics (4.B-4.C). Discussion 5.A-5.B then interprets discrimination and reliability without attributing "weak separability" to cases that would instead indicate systematic misranking under the diagnostic principle identified by the reviewer.

11. “Overinterpretation of Loss Curves: You infer from learning curves that: “EfficientNet-B4’s pretrained features already encode rich semantic cues applicable to political imagery.” This is not directly supported by loss trajectories alone. Loss convergence under frozen-backbone training shows:Optimization stability, Head-level separability. It does not demonstrate semantic alignment or task understanding.”

I removed any claim that learning curves demonstrate “semantic understanding” or pretrained political alignment, and I avoided using loss convergence as evidence of task understanding. In the revised framing, optimization behavior is treated only as a training diagnostic rather than a basis for semantic conclusions. Moreover, the Methods section explicitly states that all reported results are obtained via full end-to-end fine-tuning (3.B), and does not present frozen-backbone/head-only training as the primary evidence base; consequently, the Results section does not argue from head-only learning curves to semantic conclusions (4.A-4.Q). Any remaining discussion of training dynamics is confined to appropriately cautious language in Discussion 5.G. It emphasises that convergence indicates optimization stability and separability under the training regime rather than mechanistic understanding of political content.

12. “Strong Language in Results Section Is Inappropriate. Results sections should report, not argue. Phrases such as:“This confirms that…”“This demonstrates that…”“The model understands…”“Rendering it unusable”, belong in the Discussion, not the Results. Their presence weakens scientific neutrality.”

I fully rewrote the Results section to maintain scientific neutrality by reporting measured quantities, experimental conditions, and descriptive comparisons without argumentative or confirmatory phrasing, and interpretive statements are reserved for the Discussion. Concretely, Results 4.A-4.Q uses descriptive language (“reports,” “shows,” “varies,” “is associated with,” “is consistent with”) and ties each claim to a figure/table and an explicit evaluation definition (Methods 3.G), avoiding causal or confirmatory wording. Interpretive synthesis, such as how to view stochastic inference as a control for noise-induced smoothing, and how to treat uncertainty as a decision-support signal, is consolidated in Discussion 5.B-5.F, so that the Results section remains a neutral accounting of empirical observations rather than an argumentative narrative.

Thank you for addressing my concerns so comprehensively. Much appreciated. The manuscript is much improved and adds value as a diagnostic paper.

However, now we are left with usability concerns and/because of the conflation of confidence with uncertainty. Please address my concerns below. Point One is a test that you could perhaps perform on your model. Point two is primarily for discussion and conclusion. None of these are deal-breakers.

1.The paper never shows:Uncertainty conditioned on confidence. e.g., error rate at fixed confidence but varying uncertainty (see point 2 also). Without this, you cannot claim separation (of uncertainty and confidence). A clean test would have been:Fix predicted probability ≈ 0.9. Compare low-variance vs high-variance predictions. See if error rates differ.

2.Your title is misleading. It should actually state the opposite, such as "“Uncertainty estimation does not reliably indicate error in political deepfake detection” , because your paper shows that adding stochasticity (MC dropout, noise) can marginally (non-significantly) improve calibration while making uncertainty WORSE as an error signal. While this is non-intuitive, it is not of much use since now uncertainty estimates from stochastic CNNs cannot be treated as a trust signal unless uncertainty–error alignment is verified? Hence, there is no quantification of whether I should trust a high uncertainty, marginally greater true predictability signal or a low uncertainty, marginally lesser true predictability signal. Can you perhaps quantify this in some shape or form? Please also discuss in some depth in the manuscript.

I deeply thank the reviewer for the feedback on the missing pieces. I also implemented the optional test for Point One. Below, I reproduce each comment verbatim and describe precisely how and where it has been addressed in the revised manuscript.

13. *The paper never shows:Uncertainty conditioned on confidence. e.g., error rate at fixed confidence but varying uncertainty (see point 2 also). Without this, you cannot claim separation (of uncertainty and confidence). A clean test would have been:Fix predicted probability ≈ 0.9. Compare low-variance vs high-variance predictions. See if error rates differ.*

In the revised manuscript, I have added an explicit confidence-conditioned uncertainty analysis that directly implements the proposed test. Specifically, in Section 4 (Results), Subsection 4.J "Uncertainty conditioned on confidence: error rates at fixed predicted probability", I restrict evaluation to predictions whose Monte Carlo predictive mean satisfies **\hat{\mu}(x)\in[0.85,0.95]**, operationalizing "predicted probability approx. 0.9". Within this fixed-confidence band, predictions are split into low- and high-uncertainty groups using a within-band median split of MC predictive variance computed with T=20 stochastic forward passes. I then separately report empirical error rates for each group using the same fixed decision threshold (t=0.5) and label semantics defined throughout the paper. I reported the resulting counts and error rates in Table 4.V, and the accompanying text explicitly states that, under fixed confidence, higher predictive variance corresponds to higher empirical error rates for both EfficientNet-B4 and ResNet-18. This analysis directly demonstrates separation between confidence and uncertainty without relying on global comparisons.

14. *Your title is misleading. It should actually state the opposite, such as '"Uncertainty estimation does not reliably indicate error in political deepfake detection" , because your paper shows that adding stochasticity (MC dropout, noise) can marginally (non-significantly) improve calibration while making uncertainty WORSE as an error signal. While this is non-intuitive, it is not of much use since now uncertainty estimates from stochastic CNNs cannot be treated as a trust signal unless uncertainty–error alignment is verified? Hence, there is no quantification of whether I should trust a high uncertainty, marginally greater true predictability signal or a low uncertainty, marginally lesser true predictability signal. Can you perhaps quantify this in some shape or form? Please also discuss in some depth in the manuscript.*

I respectfully retain the title "Uncertainty-Aware Political Deepfake Detection with Stochastic Convolutional Neural Networks", as it accurately reflects the scope and intent of the study. The manuscript does not claim that uncertainty estimates are universally reliable or that stochastic inference inherently improves trust. Rather, the work evaluates uncertainty-aware inference empirically and characterizes the specific conditions under which uncertainty estimates are operationally informative. To avoid overinterpretation, I have revised the manuscript to clarify that uncertainty should be treated as a conditional, decision-level signal, not as a global or monotonic trust metric. I explicitly clarified this in Section 5 (Discussion), particularly Subsections 5.C, in "Confidence-uncertainty disentanglement at a fixed operating confidence" and 5.D, and the last paragraph in Section 6 (Conclusions), where uncertainty is framed as a secondary risk stratification signal that refines decision-making given a fixed confidence level, rather than replacing confidence as a primary decision criterion. Under this framing, the retained title remains accurate and does not overstate the reliability or interpretability of uncertainty estimates.

Thank you for addressing my comments. However,

1.You have performed the test. Thank you. Your median split is arbitrary and T=20 stochastic passes are borderline. In view of this, please rewrite your conclusion from this test as "evidence consistent with separation", not definitive proof.

2.As mentioned, I would have liked to see some quantitative definition for the following tradeoff "Should I trust a high-uncertainty, slightly better predictor or a low-uncertainty, slightly worse one? " The manuscript still shows only slight calibration improvement, worse uncertainty–error alignment globally and a narrow conditional effect that only appears after slicing. Is it possible for you to perform any of the following empirical tests? [1] Expected risk curves with uncertainty gating, [2] Selective prediction / rejection curves, [3] Risk–coverage tradeoffs, [4] Utility-weighted decision analysis, [5] AUROC of uncertainty as an error detector at fixed confidence. [6] any other test you can think of that quantifies the trade-off? If you are unable to quantify, I still suggest

rewording the title to avoid misrepresentation of uncertainty upfront as being global and monotonic. Some suggestions are:
When and How Uncertainty Is Informative in Political Deepfake Detection
Uncertainty as a Conditional Decision Signal in Political Deepfake Detection
Evaluating Decision-Level Uncertainty in Political Deepfake Detection with Stochastic Convolutional Neural Networks
On the Conditional Use of Uncertainty in Political Deepfake Detection with Stochastic Convolutional Neural Networks
Conditional Uncertainty-Aware Political Deepfake Detection with Stochastic Convolutional Neural Networks
Decision-Level Uncertainty-Aware Political Deepfake Detection with Stochastic Convolutional Neural Networks

---

I deeply thank the reviewer for the feedback on the missing pieces. Below, I reproduce each comment verbatim and describe precisely how and where it has been addressed in the revised manuscript.

15. *You have performed the test. Thank you. Your median split is arbitrary and T=20 stochastic passes are borderline. In view of this, please rewrite your conclusion from this test as "evidence consistent with separation", not definitive proof.*

First, in Section 4 (Results), at the conclusion of the confidence-conditioned analysis (Section 4.J), I explicitly constrain the interpretation of the experiment. The revised text now states that the observed differences in error rates are specific to the chosen confidence band, the variance partitioning strategy, and the finite Monte Carlo budget, and that they should not be interpreted as evidence of global confidence-uncertainty separation across the full score distribution. I added this qualifying sentence explicitly to prevent overgeneralization.

Second, I emphasize that the confidence-conditioned analysis is reported descriptively in the Results section, without argumentative language. No claims of "disentanglement," "proof," or general separation are made in Section 4.

Third, in Section 5 (Discussion), the interpretation of this experiment is further narrowed. The relevant paragraph now explicitly characterizes the findings as evidence consistent with partial separation under this conditioning, and notes that the arbitrariness of the median split and the use of T=20 stochastic passes preclude stronger conclusions. This phrasing directly mirrors the reviewer's requested formulation and replaces earlier, stronger language.

Finally, in Section 6 (Conclusions), I summarized the confidence-conditioned findings using the same qualified framing. The conclusion reiterates that uncertainty stratifies risk only conditionally and only under explicit slicing, and that no claim of general or global confidence-uncertainty separation is made.

16. *As mentioned, I would have liked to see some quantitative definition for the following tradeoff "Should I trust a high-uncertainty, slightly better predictor or a low-uncertainty, slightly worse one? " The manuscript still shows only slight calibration improvement, worse uncertainty–error alignment globally and a narrow conditional effect that only appears after slicing. Is it possible for you to perform any of the following empirical tests? [1] Expected risk curves with uncertainty gating, [2] Selective prediction / rejection curves, [3] Risk–coverage tradeoffs, [4] Utility-weighted decision analysis, [5] AUROC of uncertainty as an error detector at fixed confidence. [6] any other test you can think of that quantifies the trade-off? If you are unable to quantify, I still suggest rewording the title to avoid misrepresentation of uncertainty upfront as being global and monotonic. Some suggestions are:*
When and How Uncertainty Is Informative in Political Deepfake Detection
Uncertainty as a Conditional Decision Signal in Political Deepfake Detection
Evaluating Decision-Level Uncertainty in Political Deepfake Detection with Stochastic Convolutional Neural Networks
On the Conditional Use of Uncertainty in Political Deepfake Detection with Stochastic

Convolutional Neural Networks
Conditional Uncertainty-Aware Political Deepfake Detection with Stochastic Convolutional Neural Networks
Decision-Level Uncertainty-Aware Political Deepfake Detection with Stochastic Convolutional Neural Networks

First, I added a new Results subsection, Section 4.K, titled "Risk-Coverage Trade-offs under Uncertainty Gating." This subsection introduces an explicit quantitative operationalization of the reviewer's question by evaluating risk-coverage behavior under uncertainty-based, confidence-based and random rejection. Risk is defined as empirical error on retained predictions, coverage as the fraction of predictions retained and performance is summarized via risk-coverage curves and AURC values.

Second, the Results text in Section 4.K is written conservatively. I explicitly note that all methods coincide at full coverage and that differences arise only when selective rejection is applied. I do not claim global superiority of uncertainty-based methods, but report ordering and magnitude of effects descriptively.

Third, in Section 5 (Discussion), Subsection 5.E, titled "Quantifying the risk-coverage trade-off under uncertainty-based rejection," I directly address the reviewer's question. The discussion now states that the trade-off cannot be resolved globally: at full coverage, differences reduce to small accuracy-calibration trade-offs, and uncertainty does not justify preferring one predictor over another. The value of uncertainty emerges only when selective rejection is permitted, and even then the effect is modest and coverage-dependent.

Fourth, in Section 6 (Conclusions), I added a dedicated paragraph explicitly resolving the reviewer's question. The conclusion now states that the choice is not between a "high-uncertainty but better" and a "low-uncertainty but worse" predictor at the model level, but between operating points that trade coverage for risk. This framing avoids any implication that uncertainty provides a global or monotonic trust signal.

Fifth, throughout the revised manuscript I made some targeted micro-edits to remove or soften language that could be interpreted as presenting uncertainty as globally informative. In particular, I have been replaced with conditional, operational language, statements implying dominance, general reliability or monotonic trust.

Finally, to ensure that the framing of the paper does not overpromise the role of uncertainty, even after the new quantitative proof, I revised the title to: "Conditional Uncertainty-Aware Political Deepfake Detection with Stochastic Convolutional Neural Networks." It now reflects the manuscript's central empirical conclusion: uncertainty is informative only under explicit conditioning and at the decision level, rather than as a universal trust metric.

You have answered my concerns in letter but not in spirit. I am not trying to nitpick; I am not yet convinced that this paper is not describing a careful negative result dressed up as a conditional insight.

As a LAST experiment:

1.Please replace the single, hand-picked conditioning choice with a systematic confidence band sweep over both axes that currently prop up the main claim. Instead of only $\hat{\mu} \in [0.85, 0.95]$, Evaluate multiple bands:[0.55, 0.65],[0.65, 0.75], [0.75, 0.85], [0.85, 0.95]. For each band: Keep sample count fixed (subsample if needed). Compute uncertainty–error separation within that band.

2.Variance partition sweep. Instead of a single median split, perform Top/bottom 20%, Top/bottom 30%, Top/bottom 40% ,Median.

Report One figure + one table:

Figure: X-axis: confidence band center, Y-axis: effect size (Δ error rate or odds ratio per σ^2). Error bars: bootstrap CI

Table: % of bands where effect is significant. Directional consistency (always higher $\sigma^2 \rightarrow$ higher error?). Sensitivity to partition choice

This answers all the questions.
The effect's presence/absence across confidence is shown.
The effect persists (or doesn't) across multiple thresholds or continuous models.
You now know whether the phenomenon is local, rare, or systematic.
Even if the effect does not hold everywhere; i.e if the result is: "Uncertainty stratifies residual risk only in the upper-confidence regime, and collapses elsewhere", that still allows you to honestly upgrade your current ambiguous result of ""Here's a narrow effect—please don't overinterpret" to a much more operable and defensible result of "Here is the map of where uncertainty helps and where it provably doesn't."
I look forward to reviewing your last iteration of this promising manuscript.

"You have answered my concerns in letter but not in spirit. I am not trying to nitpick; I am not yet convinced that this paper is not describing a careful negative result dressed up as a conditional insight.
As a LAST experiment:
17. *Please replace the single, hand-picked conditioning choice with a systematic confidence band sweep over both axes that currently prop up the main claim. Instead of only $\hat{\mu} \in [0.85, 0.95]$, Evaluate multiple bands:[0.55, 0.65],[0.65, 0.75], [0.75, 0.85], [0.85, 0.95]. For each band: Keep sample count fixed (subsample if needed). Compute uncertainty–error separation within that band.*
18. *Variance partition sweep. Instead of a single median split, perform Top/bottom 20%, Top/bottom 30%, Top/bottom 40% ,Median.*
Report One figure + one table:
Figure: X-axis: confidence band center, Y-axis: effect size (Δ error rate or odds ratio per σ^2). Error bars: bootstrap CI
Table: % of bands where effect is significant. Directional consistency (always higher $\sigma^2 \rightarrow$ higher error?). Sensitivity to partition choice
This answers all the questions.
The effect's presence/absence across confidence is shown.
The effect persists (or doesn't) across multiple thresholds or continuous models.
You now know whether the phenomenon is local, rare, or systematic.
Even if the effect does not hold everywhere; i.e if the result is: "Uncertainty stratifies residual risk only in the upper-confidence regime, and collapses elsewhere", that still allows you to honestly upgrade your current ambiguous result of ""Here's a narrow effect—please don't overinterpret" to a much more operable and defensible result of "Here is the map of where uncertainty helps and where it provably doesn't."
I look forward to reviewing your last iteration of this promising manuscript."

I deeply thank the reviewer for this careful and substantive critique. I have fully rewritten certain sections.
Specifically, I have replaced the single confidence-conditioning interval with a systematic confidence-band sweep, as requested. This new experiment is reported in Section 4.J (Confidence-band sweep with variance-partition sweep). Instead of conditioning only on $\hat{\mu} \in [0.85,0.95]$, I evaluate four disjoint confidence bands spanning the score range [0.55,0.95], namely [0.55,0.65], [0.65,0.75], [0.75,0.85] and [0.85,0.95]. For each band and each backbone, the number of samples is held fixed via subsampling, ensuring that observed differences are not driven by varying sample counts across confidence regimes. Within each band, I quantified each uncertainty-error separation using a fixed decision rule and reported it as an effect size **Δ**Err, together with bootstrap confidence intervals.
In parallel, I implemented the requested variance-partition sweep. Within each confidence band, uncertainty is no longer split using a single median threshold. Instead, I evaluated four alternative partitioning schemes. These are median split, top/bottom 40%, top/bottom 30%, and top/bottom 20%, applied to the same fixed subsample within each band. This explicitly tests sensitivity to the arbitrariness of the uncertainty threshold. The complete grid of confidence-band × variance-

partition results is reported in Table 4.J, which additionally summarizes statistical significance rates/directional consistency/ sensitivity to partition choice. Figure 4.J plots effect size as a function of confidence-band center, with bootstrap confidence intervals, exactly as requested.
Crucially, this redesigned experiment makes the presence and absence of the effect explicit. The results show that, across the lower and mid-confidence bands, uncertainty-error separation is weak and statistically indistinguishable from zero across all partition schemes. It indicates that uncertainty does not provide information beyond predicted confidence in these regimes. In contrast, a consistent positive separation emerges only in the highest-confidence band, where higher predictive variance is associated with higher empirical error. Even there, the effect remains modest in magnitude and becomes statistically detectable primarily under more extreme variance partitions.
I have then rewritten Discussion Section 5.C (Operational scope of uncertainty as a conditional decision signal) to reflect this full characterization rather than a single conditional demonstration. The discussion now explicitly distinguishes between regimes where uncertainty does not stratify risk (low and mid confidence) and the regime where it does (high confidence), and it explains why strong global uncertainty-error metrics can coexist with localized conditional effects. I have fully revised the discussion section and removed redundant restatements. The section is structured to convey one negative result, one positive regime, and one operational implication, thereby avoiding overinterpretation while remaining precise.
Finally, I have fully rewritten the Conclusions (Section 6) to align with these revised findings. The conclusion no longer presents uncertainty as a generally reliable trust signal, nor does it hedge around a narrow conditional effect. Instead, it explicitly states that uncertainty provides decision-relevant information only under restricted operating conditions, specifically, by stratifying residual risk among already high-confidence predictions, and that outside these conditions it offers little additional value beyond confidence. In this way, the revised manuscript replaces an ambiguous conditional claim with an explicit map of where uncertainty helps and where it probably does not.
I believe these changes address the reviewer's concern in substance. **Thank *you* again!**

Thank you for addressing my comments. Accepted.
I will promote to copyediting.
With your permission, I would like to add the following two points under the limitations section:
The metadata-based political filtering pipeline may introduce prompt- or generator-specific confounds. There is a risk that political relevance is entangled with generation artifacts rather than semantic political content in the image. Given the near-saturated discrimination performance (AUC ≈ 0.99), it remains unclear whether the localization of uncertainty–error stratification to the high-confidence regime is intrinsic or an artifact of benchmark ease. Consequently, the operational scope of uncertainty-aware inference demonstrated here should be interpreted as dataset- and regime-specific rather than as a general property of deepfake detectors.