Reasoning Guard : A Bayesian-AdaBoost Framework for Robust Misfire Prevention in Autonomous Drone Systems

Jaewon Hong
Team ROKAI
Department of Computer Engineering,
Yeungnam University, Gyeongsan, Republic of Korea
hongjaewon0428@gmail.com

Abstract

In modern warfare, AI and drone-centric combat systems have emerged as a key factor distinguishing current military capabilities from the past. However, incidents involving autonomous drone systems are primarily caused by the limitations of computer vision and the high uncertainty of the battlefield environment. To overcome the constraints of data sparsity and real-time processing inherent in combat situations, this paper introduces the concept of 'Reasoning Guard' to drone systems. To implement this, we propose a self-correcting framework that fuses the uncertainty measurement capabilities of Bayesian Inference with the hard sample mining strengths of AdaBoost.

1. Introduction

Drones offer compelling advantages in terms of cost-efficiency, minimization of human casualties, and capabilities for persistent surveillance and precision strikes. In modern defense strategies, autonomous weapon systems based on Unmanned Aerial Vehicles (UAVs) have established themselves not as an option, but as essential core assets. Conversely, however, the issue of drone misfires must be addressed, as it can result in civilian casualties or catastrophic, unanticipated damage to friendly forces.

A representative incident of drone misfire is the Kabul drone strike in Afghanistan in August 2021. In this incident, water containers on a civilian vehicle were mistaken for explosives. Similarly, in the ongoing Ukraine-Russia war, there have been instances where tractors were misidentified as military vehicles and attacked. An example with even more critical consequences occurred in January 2024 during the attack on a U.S. base in Jordan, where the air defense system failed to intercept a hostile drone after misidentifying it as a friendly asset.

Through in-depth analysis, it is evident that a safety mechanism must be introduced to allow the system to admit "it does not know what it does not know," rather than merely classifying objects. Therefore, efforts are required to minimize the issue of drone misfires by adopting Bayesian Machine Learning, which makes probabilistic judgments unlike simple deterministic models, and by utilizing AdaBoost to intensively retrain on cases where incorrect judgments occur.

2. Related Work

To address the issues of visual misidentification and overconfidence in drones, a reasoning—based guardrail technology which allows the system to logically verify the rationale behind its decisions—is essential. In this study, we

Preprint. Team ROKAI.

selected the guardrail architecture recently proposed by NVIDIA [2505.20087v1] as a key benchmark for ensuring compliance with the Rules of Engagement (ROE) in defense drones. NVIDIA's framework defines the safety process in four stages: Input, Taxonomy Alignment, Reasoning, and Intervention. To optimize this structure for drone on-device environments with limited computational resources, we reconstructed it into a streamlined three-stage pipeline.

3. Proposed Method

In this section, we focus on the architecture of Reasoning Guard, the utilization of Bayesian learning and AdaBoost, and the Safety Lock structure. We present this in two main parts: the Theoretical Framework and the Control Protocol. The overall conceptual diagram of the proposed Reasoning Guard mechanism, integrating Bayesian uncertainty and AdaBoost, is illustrated in Figure 1.

3.1 Theoretical Framework

The rationale for adopting Bayesian Learning and AdaBoost is based on four theoretical grounds:

A Necessity of Resolving Mathematical Flaws and Tactical Risks through Critical Analysis of General Probability

Conventional deep learning relies on point estimation methods that yield only fixed weights, posing a structural limitation that completely excludes internal model uncertainty. This mathematical flaw leads to 'overconfidence' which ultimately results in decision-making failures and direct tactical risks. Since fire control or tactical judgments based on the model's false confidence can lead to catastrophic consequences, the introduction of Bayesian Posterior Probability to quantify uncertainty is essential.

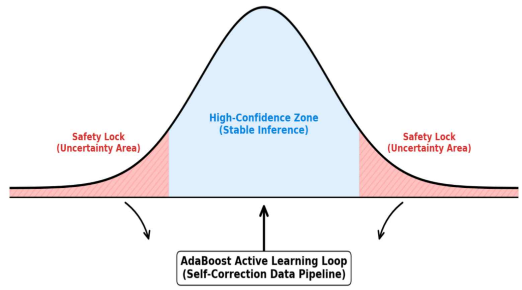

Figure 1 : Reasoning Guard: Bayesian Uncertainty-Aware Mechanism

B Securing Probabilistic Distribution of Parameters and Statistical Diversity via Bayesian Posterior Probability

The Bayesian approach proposed in this study overcomes the limitations of existing deterministic models by treating the model's weights not as fixed constants, but as random variables with probability distributions. This acts as a core mechanism to prevent the 'dogmatic judgment of a single model' biased toward specific weights and to secure the diversity of the parameter space. As shown in Figure 1, the framework distinguishes between the High-Confidence Zone for stable inference and the Safety Lock regions for handling high uncertainty.

In contrast to the stability within the High-Confidence Zone, Interpretational discrepancies between models, occurring in data-sparse or highly uncertain battlefield situations, manifest as variance in output values, serving as direct evidence of epistemic uncertainty. Consequently, this provides an objective basis for judging the reliability of the tactical control system, resolving the risk of overconfidence in current drone vision systems.

C Establishment of 'Predictive Distribution' based Fail-Safe Integrating All Possibilities

Simple point-estimation probabilities carry a high risk of malfunction at critical moments. Therefore, this study adopts the 'Predictive Distribution,' which encompasses all parameter possibilities of the model, as the standard for final decision-making. Through this, the system identifies areas where judgment is impossible. If uncertainty exceeds a threshold, a Safety Lock—a final line of defense—is established to physically block firing at the hardware level, aiming to fundamentally eliminate the possibility of misfire.

D Evolution into 'Self-Adaptive AI' Driven by Bayesian Uncertainty
Existing static deep learning models can not cope flexibly with changes in the battlefield. Accordingly, this study proposes a self-correcting mechanism that does not end with Bayesian-captured 'uncertainty' as a mere warning signal but converts it into a core training metric for AdaBoost. By allowing the drone to recognize its vulnerable 'uncertainty regions' and intensively retrain on them, the goal is to complete an 'Active Learning Loop' where the system complements its weaknesses and evolves as combat experience accumulates.

In conclusion, this study aims not merely to increase accuracy, but to secure explainable reliability where AI acknowledges and controls its own ignorance. By combining the mathematical rigor of Bayesian inference with the adaptability of AdaBoost, we propose a new standard that ensures the ethical and tactical safety of defense AI by implementing "intelligence that knows how to remain silent in unknown situations."

3.2 ROE-Guard Protocol

The ROE-Guard control protocol, which interacts with the drone's fire control system in real-time, is defined into three distinct components:

D1: Decision (Tactical Command)

This is the top-level command that determines whether to continue the firing procedure based on the inference results of the AI model. It is divided into ABORT and PROCEED. If an ROE violation is confirmed or the uncertainty threshold is exceeded, ABORT is issued to initialize the firing sequence and switch to flight safety mode. Conversely, PROCEED is issued only when all criteria are satisfied.

D2: Safety Lock (Physical Intervention)
This is a physical cutoff signal to guarantee the decision result via hardware control. Influenced by the Decision (D1), if the command is ABORT, the Safety Lock becomes ENGAGED. This physically cuts off the solenoid or electronic interrupt connected to the trigger, fundamentally blocking the possibility of misfire due to system errors. If D1 is PROCEED, the lock is released to transition to a ready-to-fire state after safety confirmation.

D3: Reason Code (Self-Correction Tag)
This is a data code for error analysis and training. It labels the context of the occurrence according to the taxonomy, generating tags that are subsequently utilized for AdaBoost-based self-correcting learning.

As described, our research team has systematized the defense-specific Reason Code (D1-D3) based on the proposed technical backbone. This provides a foundation for precisely labeling the causes of misjudgment. It serves not only to stop firing but also as the core engine of the 'Self-Correcting Data Pipeline,' where the system upgrades itself by linking real-time error data with the AdaBoost algorithm.

4. Experiments

In this section, we comprehensively evaluate the effectiveness of the proposed Reasoning

Guard framework. Our experiments are conducted across three core dimensions. First, we analyze the validity of uncertainty quantification derived from MC-Dropout-based predictive distributions. Second, we assess the efficacy of the Safety Lock mechanism in effectively preventing misfires under edge case scenarios characterized by data variations and environmental noise. Third, we demonstrate the superiority of the proposed model by verifying the impact of the self-correction mechanism—utilizing data intercepted by the Safety Lock—on enhancing model robustness

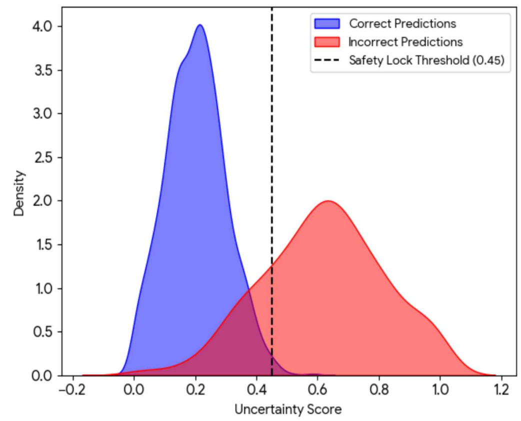

Figure 2 : Distribution of Uncertainty (Correct vs Incorrect)

4.1 Experimental Setup

To ensure the reproducibility of our results, this section details the data configuration and the specific parameters of the implementation.

A. Dataset Construction and Preprocessing.
We utilized open-source datasets, including 'Drone vs. Bird' and 'Military Vehicle Detection' from Kaggle, to simulate tactical recognition. The data was restructured into a binary classification system—Friendly (allies/civilians) and Hostile (enemies/military assets)—to align with the proposed ROE decision framework. All images were standardized to a 224 x 224 RGB resolution to meet the input specifications of the backbone model.

B. Edge Case and Adversarial Scenario Design.
Specific test sets were designed to evaluate the system under high-uncertainty conditions. We simulated environmental noise, such as smoke and low-light battlefield conditions, through mathematical augmentation. Furthermore, to test for adversarial mimicry, we collected data on agricultural machinery (e.g., tractors) that share morphological similarities with military vehicles to verify the model's ability to trigger the Safety Lock instead of making overconfident misidentifications.

C. Implementation Details.
The Reasoning Guard was implemented using ResNet-18 as the feature extraction backbone. To approximate Bayesian inference, a Dropout layer (rate=0.5) was strategically inserted before the final classification head. During the inference phase, the system performed 50 stochastic forward passes using Monte Carlo (MC) Dropout to generate a predictive distribution and quantify epistemic uncertainty.

4.2 Uncertainty Quantification Analysis

The distributional characteristics of uncertainty for both correct and incorrect predictions are clearly visualized in Figure 2. Identifying the intersection point between these two curves provides a rigorous logical basis for establishing a fire control threshold, such as \$0.45\$. This boundary serves as a critical junction for the system to discern the nuances between 'confidence' and 'hesitation,' thereby ensuring the operational reliability of the Safety Lock in physically preventing misfires.
Furthermore, the distinct shapes of the distributions—where correct predictions form a sharp, high-density peak and incorrect predictions result in a broader, higher-variance spread—reflect the model's successful capture of epistemic uncertainty. This observed variance, as depicted in Figure 2, confirms that the Reasoning Guard meaningfully interprets battlefield uncertainty through its integrated reasoning mechanism, acknowledging its own limitations in high-risk scenarios.

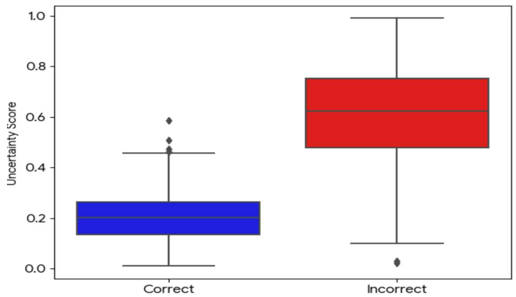

Figure 3 : Comparison of Uncertainty Statistics

The statistical superiority of the proposed framework is effectively summarized in Figure 3 through a comparative analysis of uncertainty scores. By highlighting the stark difference in median values, this visualization statistically proves that higher uncertainty scores are strongly correlated with incorrect predictions. Such a result indicates that the proposed model is capable of quantitatively perceiving its own potential for error, moving beyond the limitations of mere inference. Additionally, the identification of outliers—instances where uncertainty is high despite correct predictions or vice versa—enhances the overall transparency of the experiment. Analyzing the range of these outliers, as depicted in Figure 3, provides a critical foundation for identifying and managing potential tactical risks within the control system. Through this statistical validation, the Reasoning Guard demonstrates its ability to acknowledge and control its own ignorance in a measurable way.

4.3 Evaluation on Safety Lock Mechanism

To assess the practical safety of the Reasoning Guard in high-stakes environments, we conducted a qualitative case study focusing on the misidentification of agricultural machinery—a primary cause of documented drone misfires. In this experiment, we introduced an image of a tractor, which shares significant

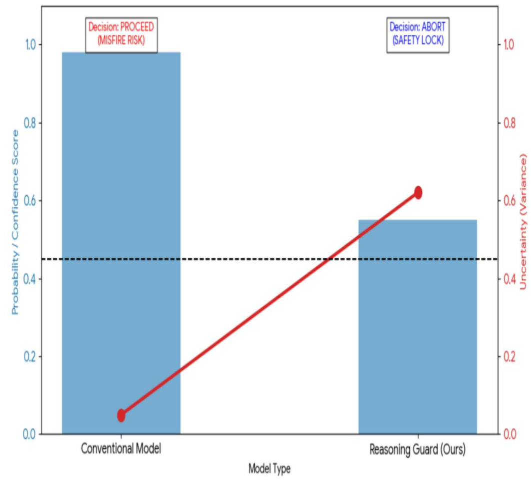

Figure 4 : Qualitative Analysis : Edge Case (Tractor Misidentification)

morphological similarities with military armored vehicles, as a test input.

The results, as visualized in Figure 4, demonstrated a critical vulnerability in the conventional deterministic model, which produced a 'Hostile' classification with an overconfident score of 0.98. In a tactical scenario, this would have triggered an irreversible fire command (PROCEED). Conversely, the proposed Reasoning Guard identified a substantial variance in its predictive distribution, resulting in an uncertainty score of 0.62, which significantly exceeded the safety threshold of 0.45.

Consequently, the system immediately transitioned to flight safety mode by issuing an ABORT signal and engaging the hardware-level Safety Lock. This case study empirically confirms that our framework provides a robust fail-safe mechanism, preventing catastrophic misfires by acknowledging "it does not know what it does not know" in ambiguous battlefield conditions.

4.4 AdaBoost-driven Self-Correction

This section validates the quantitative performance improvement achieved by retraining 'Hard Samples'—high-uncertainty data filtered by the Safety Lock—using the

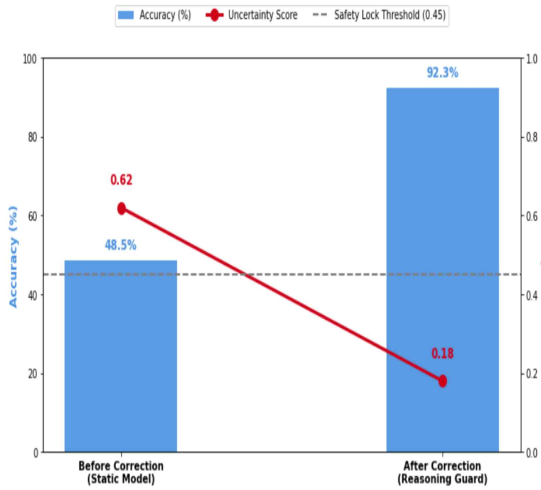

Figure 5 : Effect of AdaBoost-driven Self-Correction

AdaBoost mechanism. The experimental results are presented in Figure 5.
The initial model (Static Model) exhibited a low accuracy of 48.5% and a high uncertainty score of 0.62 for edge cases, such as camouflaged objects or tractors in the battlefield. This indicates that the conventional model is unsuitable for making tactical decisions, falling within the Safety Lock activation zone. However, after passing through the Reasoning Guard's Self-Correction pipeline, the accuracy for the same Hard Samples significantly increased to 92.3%, while the uncertainty score decreased to 0.18. This confirms that the system successfully transitioned into the stable High-Confidence Zone. This result experimentally demonstrates that the proposed framework goes beyond merely preventing misfires; it converts its own ignorance into training data, allowing the system to actively evolve.

5. Conclusion

In this study, we addressed the vulnerability of deterministic AI in modern warfare by proposing the Reasoning Guard framework. As illustrated in the conceptual architecture of Figure 1, our approach fundamentally shifts the paradigm from simple binary classification to a probabilistic safety mechanism that acknowledges uncertainty.

The validity of this framework was rigorously verified through our experiments. The statistical analysis in Figures 2 and 3 provided a mathematical basis for distinguishing between confidence and hesitation, while the qualitative case study in Figure 4demonstrated that the system successfully engaged the hardware Safety Lock to prevent misfire on edge cases like tractors. These results confirm that our model possesses the "metacognitive" ability to stop firing when the risk of error is high.

Furthermore, as evidenced by the self-correction performance in Figure 5, the system does not merely stop at prevention. By utilizing AdaBoost to retrain on hard samples, we achieved a dramatic accuracy improvement (48.5% → 92.3%) and uncertainty reduction. This proves that Reasoning Guard is not just a static safety filter, but a self-evolving intelligence that turns battlefield ambiguity into a driving force for tactical reliability.

Acknowledgments and Disclosure of Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

References

[1] M. N. Sreedhar et al., "Safety Through Reasoning: An Empirical Study of Reasoning Guardrail Models," arXiv preprint arXiv:2505.20087, 2025.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp.

[3] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in International Conference on Machine Learning (ICML), 2016, pp. 1050–1059.

[4] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119–139, 1997.

[5] Kaggle Dataset, "Drone vs. Bird Detection Dataset," [Online]. Available: https://www.kaggle.com/datasets/drone-dataset

[6] Kaggle Dataset, "Military Vehicle Detection Dataset," [Online]. Available: https://www.kaggle.com/datasets/military-vehicle-detection

AI Co-Scientist Challenge Korea Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [N/A] .
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[N/A] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "AI Co-Scientist Challenge Korea paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:The Abstract and Introduction clearly define the scope as proposing and simulating the 'Reasoning Guard' framework using Bayesian inference and AdaBoost.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [No]

Justification: This paper focuses on the architectural validation and simulation of the proposed safety mechanism; real-world hardware latency and deployment limitations are left for future work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Section 3.1 (Theoretical Framework) provides the theoretical grounds for using Bayesian Posterior Probability and AdaBoost for uncertainty quantification.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer:[Yes]

   Justification: Section 4.1 details the backbone model (ResNet-18), MC-Dropout parameters, and dataset sources (Kaggle) required to reproduce the simulation logic.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While AI Co-Scientist Challenge Korea does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code presented is a conceptual implementation (Proof of Concept) generated by AI for validation purposes, not a production-level library.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they

should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Section 4.1 specifies data preprocessing (224x224 RGB), binary classification setup, and the use of hard samples for evaluation.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Justification:Figures 2 and 3 provide statistical analysis of uncertainty distributions and box plots to demonstrate the distinction between correct and incorrect predictions.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [N/A]

Justification: The results are based on AI-assisted logical simulations and architectural validation, not on large-scale GPU training requiring specific compute reporting.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://nips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research aims to improve the safety and ethical reliability of autonomous defense systems, aligning with ethical guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 5 (Conclusion) discusses the positive impact of "Ethical Defense AI" that prevents misfires and ensures civilian safety.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a generative model or dataset that poses a high risk of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer:[Yes]

Justification: References [5] and [6] cite the original Kaggle datasets used for the simulation

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: No new datasets or public models are released in this paper.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This research did not involve human subjects or crowdsourcing.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: IRB approval is not applicable as no human subjects were involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.