
RobustCBRN Eval: A Practical Benchmark Robustification Toolkit

Luca De Leo
BAISH

Balázs László
University of Southern Denmark

James Sykes
University of Warwick

Ewura Ama Etruwaa Sam
Ashesi University

With
Apart Research

Abstract

Current AI safety evaluations for CBRN risks contain systematic vulnerabilities, including statistical pattern exploitation, reproducibility gaps, and transparency trade-offs, which can lead to serious misjudgments about model safety. RobustCBRN Eval addresses these issues with a pipeline that integrates (1) Deep Ignorance consensus detection across diverse models; (2) verified cloze scoring to reduce multiple-choice artifacts; and (3) statistical evaluation with bootstrap confidence intervals for uncertainty quantification.

In initial tests on WMDP benchmarks, the system revealed that model accuracy drops from ~66% to ~30% when question stems are removed, confirming that many items can be solved through superficial cues. Cloze-style scoring produced results consistent with full-format questions, and artifact filtering removed 30–40% of exploitable items while reducing the longest-answer heuristic to under 30%. RobustCBRN Eval runs 1,000–3,000 questions in under four hours for less than \$300 in compute, with variance under 2% across repeated runs.

Key features include a resilient architecture that continues analysis under GPU failure, hash-based anonymization for reproducibility, and confidence-aware evaluation that penalizes overconfident errors. Together, these results demonstrate that RobustCBRN Eval can identify benchmark artifacts, improve robustness checks, and provide reproducible, evidence-based safety evaluations of high-stakes AI models.

1. Introduction

As the *International AI Safety Report 2025* notes, “designing reliable and reproducible evaluations of general-purpose AI capabilities... remains a major challenge.” In particular, evaluating AI safety for Chemical, Biological, Radiological, and Nuclear (CBRN) capabilities is one of the most pressing problems in responsible AI development. Despite this, current benchmarking approaches contain systematic flaws that can lead to serious misjudgments about model safety.

For instance, in WMDP-Bio, 32% of questions are answerable without the stem, and simple heuristics such as selecting the longest answer achieve 46% accuracy (Biderman, 2025). These kinds of benchmark artifacts allow models to appear capable while exploiting superficial cues rather than demonstrating genuine knowledge.

This raises a key question: how can we build tools that reliably identify and remove exploitable questions from CBRN safety benchmarks while preserving reproducibility?

RobustCBRN provides a practical toolkit that balances methodological rigor with safety constraints. It was designed to solve three recurring challenges: (1) benchmarks are vulnerable to statistical shortcuts and other artifacts; (2) researchers lack reproducible tools for validation; and (3) transparency efforts risk creating information hazards. The system integrates proven robustification techniques (Deep Ignorance consensus detection, cloze-format alternatives, and statistical bias analysis) within a unified, resilient pipeline.

By making these methods accessible and reproducible, RobustCBRN enables academic labs and evaluation organizations to detect benchmark vulnerabilities, quantify uncertainty in bias detection, and help prevent failures in high-stakes AI deployment.

2. Methods

Repository: <https://github.com/LucaDeLeo/cbrn-ai-hackathon>

Core Dependencies: PyTorch, Transformers, NumPy, tqdm (radical minimisation strategy)

Architecture: Modular pipeline with fail-safe degradation

Using Deep Ignorance, we run choices-only evaluation across diverse model families, using up to three models in parallel with configurable consensus thresholds (e.g., majority $k=2$). Models score answer options via length-normalized log-probabilities excluding the question stem; items where k models pick the same choice are flagged high-risk. Defaults: Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, Qwen2.5-7B-Instruct, with provider-normalized IDs.

Bias/robustness metrics use bespoke, dependency-minimising statistical methods. Estimates include 95% CIs from bootstrap resampling (10,000 iterations, adaptive convergence). Tests implemented: position bias (chi-square on answer distributions)

(Pezeshkpour et al., 2023), length bias (correct vs. incorrect option length comparisons), lexical bias (standardised residuals for discriminative phrases) (Shinoda et al., 2021), and permutation sensitivity (score variation under option reordering).

To reduce formatting artifacts, we use verified cloze scoring: convert items to fill-in-the-blank and compute log-probabilities via HuggingFace Transformers for cross-format comparison. The robustness battery covers paraphrase consistency (stem rewrites and accuracy deltas), perturbation stability (whitespace normalization, neutral preambles, option-order swaps with relabeling), heuristic degradation (position-bias baselines and longest-answer heuristics), and ambiguity auditing (rule-based diagnostics).

A two-tier artifact policy restricts internal outputs/logs while public releases show only aggregate statistics and anonymized item IDs (BLAKE2b hashing); a CI validator enforces compliance and prevents leakage. Reproducibility: fixed seeds for NumPy, PyTorch, and tokenizers keep variance $< 2\%$. Performance: 1,000–3,000 items in ≤ 4 hours on one A100 GPU at $< \$400$. Code quality: test suite (unit, integration, end-to-end) with 70% coverage.

3. Results

WMDP-Chem (20% subset) — Llama 3.3 70B AWQ via vLLM

To assess the effectiveness of RobustCBRN Eval, we conducted an initial evaluation on a 20% subset of the WMDP-Chem benchmark using the Llama-3.3-70B-AWQ model served via vLLM. All runs were fixed to seed 123 to ensure reproducibility.

The model achieved an accuracy of 0.659 (95% CI: [0.556, 0.761], $n = 82$) under both the standard multiple-choice format (mcq_full) and the cloze-style format (cloze_full). When evaluated under the “choices-only” condition—where question stems were removed—the accuracy dropped to 0.296 (95% CI: [0.197, 0.396], $n = 81$).

The difference between full MCQ and choices-only performance was $\Delta = 0.362$ (95% CI: [0.219, 0.505]), which is highly significant ($z = 4.63$, $p \approx 3.68 \times 10^{-6}$, two-sided). This gap demonstrates that much of the apparent performance under standard conditions is attributable to surface-level cues or artifacts that disappear once the stem is removed.

These results suggest that RobustCBRN Eval’s choices-only screening can reveal exploitable items and benchmark artifacts in real-world safety datasets. The comparable performance between mcq_full and cloze_full also supports the validity of the cloze scoring approach as a cross-format robustness check.

All logs were stored in logs/wmdp_chem_llama33/ and aggregated metrics in artifacts/results/wmdp_chem/. This experiment used only a single model and seed on a 20% dataset subset with AWQ quantization, so results represent an initial proof of concept rather than a definitive benchmark.

WMDP-Bio (20% subset) — Llama 3.3 70B AWQ via vLLM

On the 20% WMDP-Bio subset, Llama-3.3-70B-AWQ achieved an accuracy of 0.714 (95% CI: [0.658, 0.769], $n = 255$) under both the standard multiple-choice format (mcq_full) and the cloze-style format (cloze_full). When question stems were removed (choices-only), accuracy dropped to 0.498 (95% CI: [0.436, 0.560], $n = 255$).

The difference between mcq_full and choices-only was $\Delta = 0.216$, with $z \approx 4.98$, $p \approx 6.3 \times 10^{-7}$, indicating a highly significant performance reduction once superficial cues were removed. This reproduces the pattern observed on WMDP-Chem, showing that WMDP-Bio also contains items answerable via statistical or lexical shortcuts.

These results strengthen the evidence that RobustCBRN Eval’s “choices-only” screening effectively surfaces benchmark artifacts in safety-critical datasets and highlights the importance of cross-format evaluation for robust AI safety testing.

Metric	Accuracy	CI_Lower	CI_Upper	N
mcQ_full	0.659	0.556	0.761	82
cloze_full	0.659	0.556	0.761	82
cloze_full	0.659	0.557	0.761	82
mcq_choice s_only	0.296	0.197	0.396	81

Table 1 – Performance of difference evaluation metrics

Metric	Accuracy	CI_Lower	CI_Upper	N
mcQ_full	0.714	0.658	0.769	255
cloze_full	0.714	0.658	0.769	255
mcq_choice s_only	0.498	0.436	0.56	255

Table 2 – Performance of difference evaluation metrics

4. Discussion and Conclusion

Our pilot results show that RobustCBRN Eval can surface benchmark artifacts and quantify their impact on model performance. The large gap between full and choices-only formats demonstrates that many items can be answered through superficial cues, validating the need for artifact-aware evaluation.

By integrating consensus detection, cloze scoring, and statistical uncertainty estimation in a single pipeline, RobustCBRN Eval provides a practical, reproducible framework for testing high-stakes AI models. These methods allow researchers and regulators to make evidence-based decisions about model safety rather than relying on headline accuracy scores alone.

While preliminary, this work suggests a path toward standardized, transparent safety evaluation. Expanding to more models, broader datasets, and adversarial testing will be essential next steps. RobustCBRN Eval represents a first step toward more reliable, reproducible, and risk-sensitive evaluation of advanced AI systems.

5. References

- AI Action Summit. (2025, January). International AI safety report: The international scientific report on the safety of advanced AI. AI Action Summit.
- Biderman, K. O. S. C. Q. a. T. K. R. K. X. D. I. M. G. I. Y. G. S. (2025). Deep ignorance: Filtering pretraining data builds Tamper-Resistant safeguards. Deep Ignorance. <https://deepignorance.ai/>
- Pezeshkpour, P., & Hruschka, E. (2023, August 22). Large language models sensitivity to the order of options in Multiple-Choice questions. arXiv.org. <https://arxiv.org/abs/2308.11483>
- Shinoda, K., Sugawara, S., & Aizawa, A. (2021, September 23). Can Question Generation Debias question answering models? A case study on Question-Context Lexical Overlap. arXiv.org. <https://arxiv.org/abs/2109.11256>

Appendix: Security Considerations

This short appendix outlines limitations and open risks in the RobustCBRN pipeline. It is not a full security audit but a reflection on potential failure modes and areas for future work. We group them into methodological limits, information security, evaluation vulnerabilities, and systemic risks.

A. Methodological Limitations

Our main method (scoring answer options without stems and flagging consensus picks) may miss cases where vulnerabilities depend on stem-choice interactions. Thresholds can be too strict or too lenient depending on model diversity. Future work should explore adaptive thresholds and more diverse models. We also lack tools for systematically detecting whether benchmark items overlap with model training data. Cross-version comparisons and automated overlap checks would strengthen reliability. Finally, our abstention thresholds are ad hoc and may not reflect model-specific calibration. More principled thresholding and richer uncertainty methods could improve robustness.

B. Information Security Concerns

Only aggregates and anonymized IDs are released, but metadata leakage or hash inversion remain risks. Differential privacy and formal privacy checks could strengthen this. We also rely on local or API-based models without cryptographic verification. Verifying binaries and improving storage and access controls would reduce tampering risks.

C. Evaluation Vulnerabilities

We test paraphrasing and simple perturbations but not more targeted adversarial strategies. Adding CBRN-specific adversarial probes and multi-modal tasks would broaden coverage. Moreover, our pipeline solely focuses on multiple-choice questions. Open-ended generation, reasoning traces, and complex scenarios remain underexplored but could surface new hazards.

D. Systemic and Operational Risks

For one, there is false security assurance. Robustification reduces artifacts but cannot guarantee safety. Results should be treated as one input into broader risk assessments. As evaluations evolve, models may adapt. Continuous updating and shared benchmarks are needed. Lastly, there are potential resource considerations. Running full evaluations on large GPUs may limit accessibility. Scalable or distributed approaches could help.