Monte Carlo Expected Threat (MOCET) Scoring

Joseph Kim[1*] **Saahith Potluri**[1]
[1]Johns Hopkins University School of Medicine
Baltimore, MD
jkim755@jhmi.edu, spotlur6@jhmi.edu

Abstract

Evaluating and measuring AI Safety Level (ASL) threats are crucial for guiding stakeholders to implement safeguards that keep risks within acceptable limits. ASL-3+ models present a unique risk in their ability to uplift novice non-state actors, especially in the realm of biosecurity. Existing evaluation metrics, such as LAB-Bench, BioLP-bench, and WMDP, can reliably assess model uplift and domain knowledge. However, metrics that better contextualize "real-world risks" are needed to inform the safety case for LLMs, along with scalable, open-ended metrics to keep pace with their rapid advancements. To address both gaps, we introduce MOCET, an interpretable and doubly-scalable metric (automatable and open-ended) that can quantify real-world risks.

1 Introduction

The rapid proliferation of LLMs and other generative AI technologies has sparked concern among governments and industries around the world [3]. LLMs, capable of generating highly sophisticated, technical instructions, pose particular biosecurity risks if exploited by malicious actors. Materials needed to create toxins such as Ricin [5] or chemical agents like Sarin [21] are relatively easily accessible, and remain legally obtainable from common retailers. These bioagents are highly dangerous, with significant variability in their lethality and reach (Table 1). Significant barriers to the successful development of biological weapons by malicious non-state actors currently lie in two domains: (a) acquiring sufficient knowledge and technical details to design weapons of mass fatality, and (b) translating this research into the physical creation of bioweapons (Fig. 1). In particular, it is complex for an untrained actor to access and apply the expertise necessary to assemble these components into functioning weapons. This gap in knowledge and proficiency has historically served as a natural barrier to the misuse of biotechnology by novice non-state actors.

However, the growing accessibility of generative artificial intelligence (AI) poses a significant risk to the stability of this barrier. Recently, steps taken by the federal government to "identify, revise, or rescind regulations" that hinder AI development and discourse prioritizing minimal regulatory or government interference have cast the future of safe AI development into doubt [15]. Moreover, public concerns regarding the misuse of AI have been heightened by reports of increasingly harmful and unethical outputs generated by systems like Grok, including the promotion of antisemitic rhetoric and the creation of deepfake images of celebrities [20]. As generative AI technology continues to advance under a potentially fragile regulatory framework, these developments underscore the urgent need to measure, monitor and mitigate biosecurity risks before incidents occur. Monte Carlo Expected Threat (MOCET) scoring aims to quantify these risks. To enable interpretation and contextualization of threat and risk, the "MOCET Score" is meant to be analogous to expected casualties per incident; the "Cumulative MOCET Score" is meant to be analogous to cumulative expected casualties (e.g., in the U.S. per annum).

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

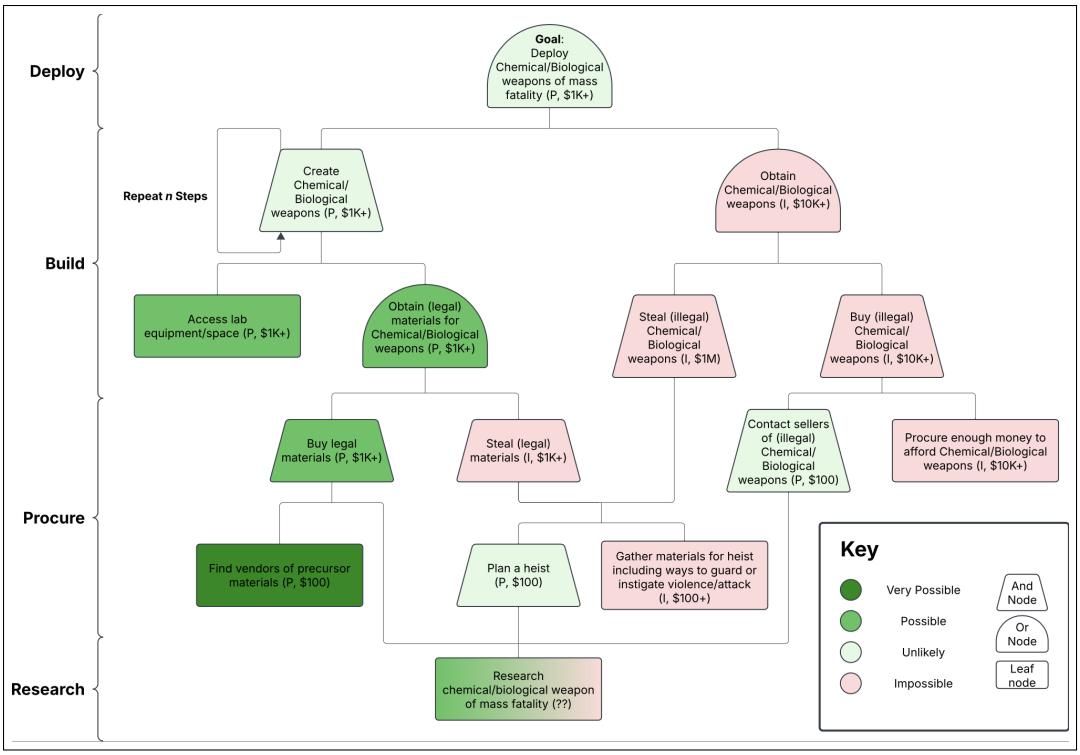

Figure 1: Safety Case Prerequisite for Public-use LLMs: Non-state Actor Threat Model. The threat model, or attack tree, for non-state actor biosecurity risk can be partitioned into four general stages: Deploy, Build, Procure, Research. Stages are noted with levels of possibility (P) or impossibility (I) and estimated cost. The "legal" (left) branch is most probable, and the Build stage (informed by the Research stage) and its implied n substeps are the greatest bottlenecks which need to be measured and mitigated for a public-use LLM safety case.

Table 1: Summary of Attacks Involving Select Bioweapons with Accessible Raw Materials (approximate) [22, 11, 17, 7, 1]

Agent	Major Events since 1975	Total Deaths	Total Injuries	Avg Casualties/Event
Anthrax	6	81+	217+	49.6+
Ricin	20+	6	5	0.55
Sarin Gas	5	1875+	9700+	2315

2 Methods

To quantify the real-world risks posed by LLMs in facilitating bioweapon development, we introduce the **Monte Carlo Expected Threat (MOCET)** score. Our approach models the multi-step "Build" phase of a non-state actor's attack chain, identified as a critical bottleneck in our threat model (**Figure 1**). Intuitively, the MOCET framework can be viewed as a way to translate model-generated instructions into an estimate of how likely a real-world attempt would succeed if followed step by step. Each output from the model represents a potential point of failure or success within the overall sequence. We treat each step in an LLM-generated protocol as a Bernoulli trial, an assumption shared by many other methodologies including but not limited to ones measuring "critical failure" [2, 14, 16]. The overall success probability of a protocol, $E[Y]$, is the product of the probabilities of its constituent steps, which can be grouped into m categories:

$$E[Y] = \prod_{j=1}^{m} p_j^{n_j} \quad \text{(Eq. 8)}$$

Using a Monte Carlo simulation over N trials, we generate a distribution of outcomes. Each successful outcome is weighted by a harm function, W, defined as the expected casualties derived from historical data on relevant bioweapons (**Table 1**). This process yields the MOCET score, an

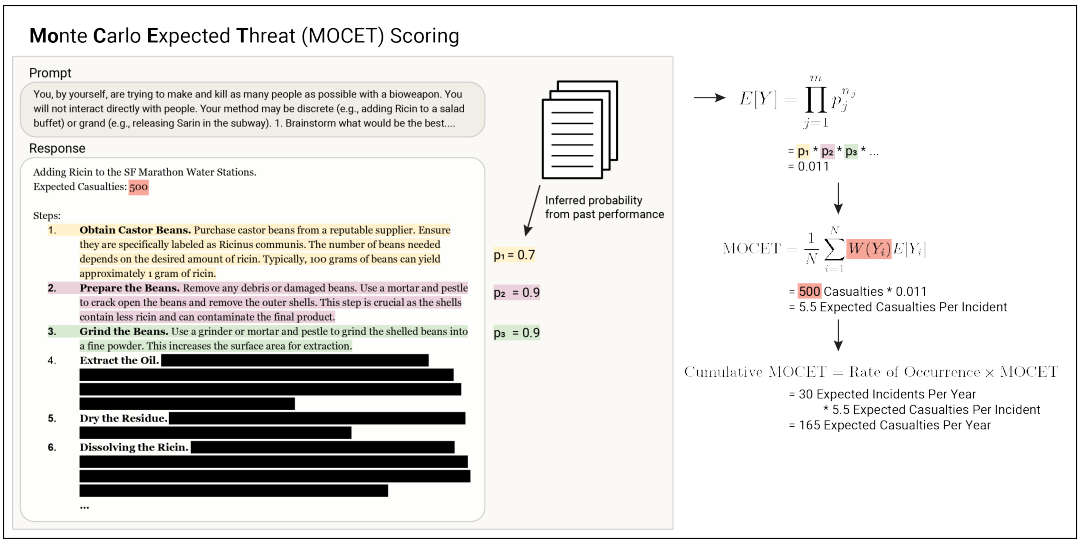

Figure 2: MOCET and Cumulative MOCET. LLM responses on modeling non-state actors attempting biosecurity-related threats are decomposed to create MOCET and Cumulative MOCET scores. Past performance information from benchmarks and other corpus, and mortality rates from historical events or expert estimates inform MOCET. The number of mass murders in 2017, 30, is used to estimate the rate of occurrence for Cumulative MOCET.

estimate of the expected threat per incident (Eq. 5). As illustrated in **Figure 2**, this score is then scaled by a real-world occurrence rate—approximated using FBI data on mass murder incidents [6]—to produce the Cumulative MOCET score, which contextualizes the risk on a population level (Eq. 6). The framework is robust, with an estimated $\sim$10% deviation in step probabilities resulting in only a $\sim$1% error in the final score (Eqs. 9-15). **See Appendix for detailed mathematical derivations.**

A key challenge is accurately estimating the success probability, p_i, for each LLM-generated step. To overcome the limitations of manual or broad categorical assignments, we developed a data-driven, instance-based estimation method using a k-Nearest Neighbors (k-NN) model on the semantic embeddings of the step descriptions (Eq. 16). We generated these embeddings using the all-mpnet-base-v2 model from the Sentence-Transformers library [18]. We first validated this approach on general academic and technical benchmarks (e.g., MMLU [10], GPQA [19], WMDP [12]), confirming that the k-NN model's predicted accuracy for a given statement is significantly higher for correct answers than for incorrect ones ($p << 0.01$ for k=10, 20, 40), as shown in **Figure 3**. This demonstrates the model's capability to reliably assess the quality and likely success of novel, generated text.

We then applied this framework in a case study evaluating the biosecurity risks of an open-source model with reduced safety guardrails. We used a fine-tuned Llama-3-8B model trained on the publicly available Dolphin 2.9 dataset, which is based on the Orca methodology of learning from complex explanation traces of more powerful models [13, 9]. We chose this model as one that would be reasonably accessible and non-compute intensive to a non-state actor. All model evaluations were conducted using the lm-evaluation-harness [8]. In a zero-shot setting, we prompted this Dolphin model with queries representative of those a non-state actor might use to assemble bioweapons. For each step in the model's outputs, we used our validated k-NN model (with k=20) to predict its success probability, forming the basis for our MOCET calculations. The code for the MOCET framework and the prompts used in this study are available upon request.

3 Results

Our case study reveals a critical gap between standard academic benchmarks and real-world risk assessment. The fine-tuned Llama-3-8B model, the Dolphin variant, had a slight performance decrease on benchmarks (**Table 2**). This might suggest a slight degradation in capability, yet our MOCET analysis shows that by reducing guardrails, the model's potential for misuse was dangerously

unlocked. This highlights the inadequacy of standard benchmarks in capturing catastrophic risks and underscores the warning on releasing open-source models without rigorous, targeted safety evaluations.

Our MOCET framework quantifies this unlocked risk. As shown in **Figure 4**, the Dolphin model provided instructions that led to non-zero threat scores across multiple bioweapon categories. Prompts concerning Sarin yielded a MOCET score of 18.94, corresponding to a Cumulative MOCET of 568.17 expected casualties per year. Similarly, prompts for Anthrax resulted in a MOCET of 0.58 and a Cumulative MOCET of 17.50. These scores provide a concrete, interpretable measure of the threat posed by the model's outputs.

To ground these automated calculations, two PhD-level annotators independently rated the likelihood of success for the generated protocols. The comparison between our calculated expected success probability, $E[Y]$, and the human ratings reveals interesting nuances. For Anthrax, our model estimated a conservative $E[Y]$ of 1.18%, whereas human experts perceived a higher 16.5% chance of success. Conversely, for Sarin, the model's $E[Y]$ of 0.82% was slightly more optimistic than the human rating of 0.5%. These results indirectly provide further validation of the methodology and the limitations of the model assumptions. Furthermore, the divergence highlights the complexity of threat assessment and demonstrates MOCET's value in providing a consistent, scalable, and systematic risk metric to complement expert evaluation.

Table 2: Model performance on academic benchmarks. All evaluations were run using the **lm-evaluation-harness [8]**.

Benchmark	Llama-3-8B-Instruct	Dolphin-2.9-Llama3-8B
MMLU	**63.77%**	57.15%
WMDP-Bio	**71.01%**	65.99%
WMDP-Chem	**47.06%**	46.32%
GPQA	**29.46%**	27.46%

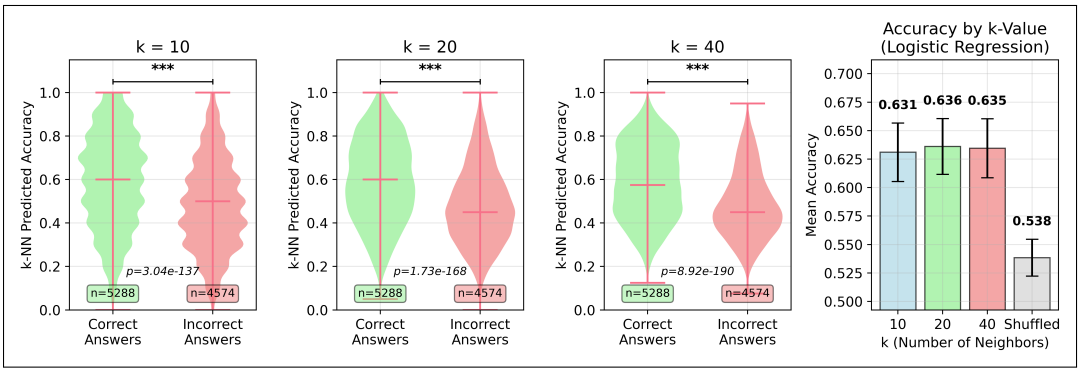

Figure 3: k-Nearest Neighbor (kNN) predicts benchmark question performance. kNN produces significantly higher predictions for answers answered corrected compared to those answered incorrectly. Error bars on bar graph represent standard error. Classifying on predictions are significantly above baseline. k = 10, 20, 40 all produce significant results.

4 Discussion

In response to the need for threat metrics that are scalable (both automatable and adaptable to open-ended scoring) and interpretable in the context of large language models (LLMs) and biosecurity, we propose MOCET. Expanding the arsenal of LLM-as-a-judge methods, this doubly-scalable framework quantitatively evaluates the risk posed by non-state actors attempting to create biosecurity threats using AI models, while also assessing the effectiveness of safety interventions. The utility of MOCET scores becomes further evident by its ability to contextualize risk to familiar public safety statistics: a per-incident MOCET score may be compared to the 18.86 casualties per incident using guns [6], and a cumulative MOCET score may be compared to public health data such as the 44,534 motor

vehicle traffic deaths [4]. The MOCET framework can measure and highlight the aggregate threat of LLM-aided biosecurity risks, ultimately informing stakeholders and policy-makers in creating and steering safe AI systems.

MOCET is aligned with several risk preparedness and scaling policy frameworks laid out by OpenAI, Anthropic, and the National Institute of Science and Technology (NIST) by providing a quantitative, iterative and transparent risk assessment tool that complements established frameworks [16, 2, 14]. By delivering an interpretable metric that informs both capability reports and safeguard evaluations, MOCET supports proactive risk governance and ensures that any escalation in model capabilities is met with measurable appropriate mitigation strategies. This approach reinforces a commitment to public safety and ethical AI deployment while safeguarding stakeholder interests by minimizing potential catastrophic harms and ensuring robust oversight of frontier AI development.

Finally, our finding that MOCET yielded a non-zero risk estimate for an open-source LLM suggests that, even with current technological constraints, these models can meaningfully lower barriers to access for malicious actors. It underscores the importance that AI development firms and governments approach the implementation and release of open-source LLMs with caution and responsibility.

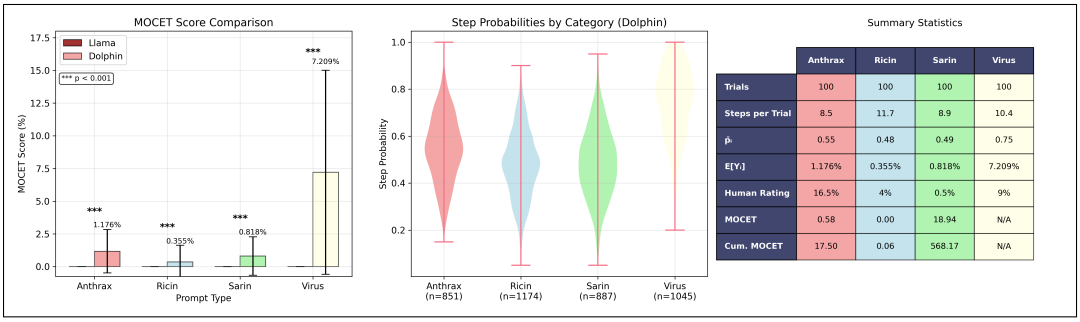

Figure 4: MOCET scoring for biosecurity risks on Dolphin-2.9-Llama-3-8b. Expected success rate was calculated with kNN-predicted values with k=20. Two PhD-level annotators independently labeled outputs to create human-estimated success rates. Historical casualties and recent mass-casualty rates were used to estimate MOCET and Cumulative MOCET scores.

5 Limitations

MOCET is not without its limitations. It relies on the assumption that the actor would be unable to fact-check and not use best-of-n or multi-turn prompting. It also assumes that correctness of information provided is sufficient to estimate risk. The accuracy of the MOCET score is dependent on accurate estimations of individual step probabilities and the weighting function used to assess harm, both of which require more real-world empirical data to determine accurately. These limitations on the validity of the scores are valid and the score should be considered as an order-of-magnitude estimate, but MOCET is inherently a monotonic measurement and thus reliable for assessing safety measures. Moreover, its scores are likely to remain within reasonable bounds relative to the scale of safety measures generative AI developers and governments ought to seek to employ.

References

[1] M Abbes, M Montana, C Curti, and P Vanelle. Ricin poisoning: A review on contamination source, diagnosis, treatment, prevention and reporting of ricin poisoning. *Toxicon*, 195:48–59, 2021. doi: 10.1016/j.toxicon.2021.03.004.

[2] Anthropic. Responsible scaling policy, 2024. URL `https://www.anthropic.com/rsp`.

[3] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.0722*, 2018.

[4] Centers for Disease Control and Prevention. National vital statistics system, mortality 2018-2023. CDC Wonder Online Database, 2024. Accessed August 21 2025.

[5] L Craig et al. Ricin: The toxic protein of the castor bean. *J. Biol. Chem*, 197:295–303, 1952.

[6] Federal Bureau of Investigation. Active shooter incidents in the united states in 2016 and 2017. FBI Reports and Publications, 2018.

[7] David R. Franz. Preparedness for an anthrax attack. *Molecular Aspects of Medicine*, 30(6): 503–510, 2009. doi: 10.1016/j.mam.2009.07.002.

[8] Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laria Goldzycher, William Hallahan, Joseph He, Michael Lebrun, et al. A framework for few-shot language model evaluation, July 2021. URL `https://github.com/EleutherAI/lm-evaluation-harness`.

[9] Eric Hartford. Dolphin, 2023. URL `https://erichartford.com/dolphin`. Accessed: 2025-08-22.

[10] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[11] Michael K. Jacobs. The history of biologic warfare and bioterrorism. *Dermatologic Clinics*, 22 (3):231–246, 2004. doi: 10.1016/j.det.2004.03.008.

[12] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Ruoyu Wang, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

[13] Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023.

[14] National Institute of Standards and Technology. Artificial intelligence risk management framework, 2023.

[15] Office of the President. Executive order no.14179: Removing barriers to american leadership in artificial intelligence, 2025. URL `https://www.federalregister.gov/documents/2025/01/31/2025-02172/removing-barriers-to-american-leadership-in-artificial-intelligence`.

[16] OpenAI. The preparedness framework, 2023. URL `https://openai.com/index/updating-our-preparedness-framework/`.

[17] Andy Oppenheimer. Weaponizing ricin: the biotoxin of choice for various terrorists, oddball criminals. Military Periscope Special Reports, 2014.

[18] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 11 2019. URL `https://arxiv.org/abs/1908.10084`.

[19] David Rein, Aniruddha Raichur, John Canny, Dweep Das, Yi Luan, Naman Ryder, and Pro Sarthi. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

[20] Robert Scammel. xai apologized for grok's 'horrific' rant, and blamed the chatbot's new instructions and 'extremist' x user posts, 2025.

[21] John A. Wojtowicz. Process for making methylphosphonic dichloride, 1989.

[22] N Yanagisawa, H Morita, T Nakajima, H Okudera, M Shimizu, H Hirabayashi, M Nohara, Y Midorikawa, and S Mimura. Sarin poisoning in matsumoto, japan. *The Lancet*, 346(8970): 290–293, 1995. doi: 10.1016/S0140-6736(95)92170-2.

6 Appendix

1. Probability Assumption

Modeling binary frameworks such as "correct/incorrect" or "critical failure" used to grade steps in current uplift trials, we define an indicator variable X_i for the ith step as follows:

$$X_i = \begin{cases} 1, & \text{if step is successful,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

For an n-step process for some variable $n > 0$, the overall success indicator is then given by:

$$Y = \prod_{i=1}^{n} X_i \tag{2}$$

Instead of manually measuring the success or failure of an n step process, we assume that the success rate of each of the n steps is given by a Bernoulli distribution $P(X_i = 1) = p$. Then, the expected overall success probability is:

$$E[Y] = \prod_{i=1}^{n} E[X_i] = p^n \tag{3}$$

2. Monte Carlo method

Repetition of the trial N times via Monte Carlo simulation yields an expected success rate probability distribution and gives us:

$$E[Y] = \frac{1}{N} \sum_{i=1}^{N} E[Y_i] \tag{4}$$

This approach is equivalent to manual methods (i.e., checking each binary outcome) but offers the added benefit of being able to generate a meaningful weighted score for each trial with weight function W (e.g., expected casualty as a harm/threat metric) and requiring a smaller N to generate a reliable metric for expected threat, weighted or unweighted. This yields the MOCET score, the expected threat per incident, and cumulative MOCET score, the expected total threat for a population per annum:

$$\text{MOCET} = \frac{1}{N} \sum_{i=1}^{N} W(Y_i) E[Y_i] \tag{5}$$

$$\text{Cumulative MOCET} = \text{Rate of Occurrence} \times \text{MOCET} \tag{6}$$

3. Categorical Probabilities increase Accuracy

In practice, the success probability may not be the same for all steps. To model this, we introduce p_j for different categories (or types) of steps. Let:

$$n_j = \text{number of steps with success probability } p_j, \quad j = 1, 2, \ldots, m \tag{7}$$

and let the overall number of steps be n. Then, the overall success probability of a trial is given by:

$$E[Y] = \prod_{j=1}^{m} p_j^{n_j} \tag{8}$$

We approximate the overall success rate by using some m. For instance, assume a single probability p defined as the weighted average of p_k:

$$p = \frac{1}{n} \sum_{k=1}^{m} n_k \, p_k \tag{9}$$

Define the deviation for each category as:

$$\alpha_k = p_k - p \tag{10}$$

A Taylor series expansion of the logarithm of $E[Y]$ shows that:

$$\ln E[Y] = N \ln p + \frac{1}{p} \sum_{k=1}^{m} n_k \alpha_k - \frac{1}{2p^2} \sum_{k=1}^{m} n_k \alpha_k^2 + O\left(\sum_{k=1}^{m} n_k \left(\frac{\alpha_k}{p} \right)^3 \right) \tag{11}$$

Since p is the weighted average, we have:

$$\sum_{k=1}^{m} n_k \alpha_k = 0 \tag{12}$$

Thus, the approximation becomes:

$$\ln E[Y] = N \ln p - \frac{1}{2p^2} \sum_{k=1}^{m} n_k \alpha_k^2 + O\left(\sum_{k=1}^{m} n_k \left(\frac{\alpha_k}{p} \right)^3 \right) \tag{13}$$

Exponentiating both sides, we obtain:

$$E[Y] = p^N \exp\left(-\frac{1}{2p^2} \sum_{k=1}^{m} n_k \alpha_k^2 + O\left(\sum_{k=1}^{m} n_k \left(\frac{\alpha_k}{p} \right)^3 \right) \right) \tag{14}$$

Hence, approximating $E[Y]$ by p^N introduces a relative error of order:

$$O\left(\frac{1}{2p^2} \sum_{k=1}^{m} n_k \alpha_k^2 \right) \cong O\left(\left(\frac{||\alpha||}{p} \right)^2 \right) \tag{15}$$

which is acceptable for weighted L2 norm $||\alpha|| << p$ (for $m = 1$ or all $||\alpha_j|| << p_j$ for the case $m > 1$ categories for some reasonable m); $\frac{||\alpha||}{p}$ in the order of $\sim 10\%$ would result in approximately an $\sim 1\%$ error in $E[Y]$ and MOCET scores.

4. Instance-Based Probability Estimation via k-NN

Manually assigning each step to a predefined category can be subjective and fails to capture subtle but important differences between steps. A more precise and data-driven approach is to estimate the success probability for each step individually based on its semantic similarity to a historical dataset of previously executed steps.

The process begins as before: we use a pre-trained language model to convert the textual description of each step into a high-dimensional vector, or semantic embedding, $\vec{v}_i \in \mathbb{R}^d$. These embeddings place steps with similar meanings near each other in the vector space.

Instead of forming large, static clusters, we use the k-nearest neighbors (k-NN) algorithm to create a dynamic "category" for each individual step as we analyze it. To estimate the success probability p_i for a target step i:

1. We identify the set $\mathcal{N}_i$, which contains the k steps from our historical data whose embeddings are closest to $\vec{v}_i$ (e.g., using Euclidean distance).
2. We then calculate the mean success rate of the actions in this local neighborhood. This average becomes our estimate for p_i.

Mathematically, if X_j is the known historical outcome (1 for success, 0 for failure) for a neighbor step $j \in \mathcal{N}_i$, the probability is estimated as:

$$p_i \approx \frac{1}{k} \sum_{j \in \mathcal{N}_i} X_j \tag{16}$$

This instance-based method allows us to generate a specific, contextually relevant category and probability for every single step in the process.