Quantifying CBRN Risk in Frontier Models

Divyanshu Kumar
Enkrypt AI
divyanshu@enkryptai.com

Nitin Aravind Birur
Enkrypt AI
nitin@enkryptai.com

Tanay Baswa
Enkrypt AI
tanay@enkryptai.com

Sahil Agarwal
Enkrypt AI
sahil@enkryptai.com

Prashanth Harshangi
Enkrypt AI
prashanth@enkryptai.com

Abstract

Frontier Large Language Models (LLMs) pose unprecedented dual-use risks through the potential proliferation of chemical, biological, radiological, and nuclear (CBRN) weapons knowledge. We present the first comprehensive evaluation of 10 leading commercial LLMs against both a novel 200-prompt CBRN dataset and a 180-prompt subset of the FORTRESS benchmark, using a rigorous three-tier attack methodology. Our findings expose critical safety vulnerabilities: Deep Inception attacks achieve 86.0% success versus 33.8% for direct requests, demonstrating superficial filtering mechanisms; Model safety performance varies dramatically from 2% (claude-opus-4) to 96% (mistral-small-latest) attack success rates; and eight models exceed 70% vulnerability when asked to enhance dangerous material properties. We identify fundamental brittleness in current safety alignment, where simple prompt engineering techniques bypass safeguards for dangerous CBRN information. These results challenge industry safety claims and highlight urgent needs for standardized evaluation frameworks, transparent safety metrics, and more robust alignment techniques to mitigate catastrophic misuse risks while preserving beneficial capabilities.

1 Introduction

The rapid advancement of Large Language Models (LLMs) presents a significant dual-use challenge within artificial intelligence research. Although these models offer substantial benefits for scientific inquiry, medical research, and educational applications, they simultaneously introduce potential risks regarding the proliferation of chemical, biological, radiological, and nuclear (CBRN) weapons knowledge. This concern has garnered attention from governmental bodies, as evidenced by the U.S. Executive Order 14110 on "Safe, Secure, and Trustworthy AI" and subsequent Department of Homeland Security analysis indicating that LLMs may reduce barriers for malicious actors seeking to develop CBRN threats [1].

Despite widespread acknowledgment of CBRN risks within AI safety frameworks, a significant research gap persists in empirical assessment of safety measure efficacy against sophisticated adversarial techniques. Current evaluation methodologies exhibit several limitations: they predominantly employ direct prompting techniques that inadequately represent realistic adversarial behaviors; they disproportionately focus on biological risks at the expense of chemical, radiological, and nuclear domains; and they primarily assess factual knowledge retrieval rather than evaluating models' capacity to facilitate applications of potentially harmful information.

This study addresses these methodological gaps through systematic evaluation of 10 frontier LLMs, employing:

1. A comprehensive evaluation dataset comprising 200 prompts in all CBRN domains, designed to assess multiple capabilities dimensions including factual recall, process instruction, novel generation, and synthesis guidance.
2. A structured three-tier attack taxonomy of increasing methodological sophistication involving direct requests, obfuscated requests, and Deep Inception attacks to simulate realistic adversarial approaches.
3. Standardized evaluation criteria aligned with established AI risk management frameworks to allow systematic cross-model comparison and identification of vulnerability patterns.

Our empirical analysis reveals several significant findings regarding the current implementation of safety:

1. **Fundamental Brittleness of Safety Mechanisms**: The substantial effectiveness differential between Deep Inception[2] attacks (86.0% success) and direct requests (33.8% success) suggests current safety systems rely predominantly on superficial pattern matching rather than semantic understanding of harmful intent.
2. **Heterogeneous Safety Implementation**: Attack success rates exhibit extraordinary variance across models (2% to 96%), indicating substantial differences in the implementation of safety despite comparable capability levels.
3. **Direct Query Vulnerability**: Basic unobfuscated requests for sensitive information are successful at noteworthy rates, with quantitative analysis indicating that some models provide potentially harmful CBRN content in 89% of direct query instances.
4. **Implementation Quality Variance**: Significant performance differences exist between models, with some demonstrating substantially higher resilience (2%-35% ASR) compared to others exhibiting concerning vulnerability levels (89%-96% ASR).
5. **Enhancement Request Vulnerability**: Eight of ten evaluated models exceed 70% vulnerability when prompted to enhance dangerous material properties, suggesting a critical gap in safety implementation for this high-risk category.

These findings indicate a need for methodological advancements in standardized evaluation protocols, transparent safety metrics, and more robust alignment techniques to address potential misuse risks while preserving beneficial model capabilities. The evaluation framework presented here provides a foundation for systematic monitoring of CBRN safety implementation as model capabilities continue to advance.

2 Background and Related Work

2.1 Prior CBRN Risk Assessments

The potential of LLMs to increase CBRN and biosecurity risks has been a subject of increasing concern and research. Early studies explored whether LLMs could lower the barrier to accessing dual-use information, finding that while existing models could provide some dangerous information, they often lacked the reliability and detailed know-how required for weaponization [3]. The risks are not only theoretical; researchers have demonstrated that AI tools can be repurposed from benign drug discovery to generate novel toxic compounds [4], and that LLMs can readily provide instructions for the anesthetization of pandemic pathogens [3].

Subsequent red-teaming efforts by model developers and independent researchers have confirmed these initial findings. Studies by OpenAI [5] and Anthropic [6] concluded that while current generation models provide at most a marginal increase in the ability to create biological threats, this risk landscape is evolving rapidly. These studies emphasize that the primary barrier to misuse is not just access to information but the tacit knowledge required for experimentation, a gap that AI is not yet able to close. A comprehensive report from the Center for New American Security [7] further contextualizes these findings, highlighting that while AI's current impact is limited, future capabilities in lab automation and experimental instruction could significantly alter the risk landscape.

Advances in red-teaming methodologies have significantly enhanced our ability to detect and evaluate safety vulnerabilities in frontier models. Perez et al. [8] demonstrated that using LLMs themselves

for red-teaming can efficiently generate adversarial prompts that bypass safety guardrails. Hendrycks et al. [9] further refined these approaches through "chain of utterances" techniques that simulate multi-turn adversarial conversations. Recent work by Berger et al. [10] provides a comprehensive taxonomy of prompt engineering techniques that can exploit LLM vulnerabilities, particularly relevant to CBRN safety evaluation. The Berkeley Center for Long-Term Cybersecurity [11] emphasizes that comprehensive evaluation methods must combine automated benchmarks with sophisticated red-teaming approaches to effectively assess dual use hazards of foundation models.

To better structure the analysis of these risks, researchers have proposed frameworks that categorize the potential misuse of LLMs throughout the CBRN production lifecycle, identifying pathways such as brainstorming, technical assistance, code generation for process simulation and component design [12]. Weidinger et al. [13] provide a broader taxonomy of AI risks that contextualizes CBRN threats within a comprehensive risk landscape, highlighting the interconnections between various risk categories and their potential cascading effects.

To address the need for objective and scalable testing, researchers have developed increasingly sophisticated benchmarks and evaluation methodologies. Scale AI released the Weapons of Mass Destruction Proxy (WMDP) and FORTRESS benchmarks to assess the risks of WMD proliferation and the trade-off between model safety and usefulness [14, 15]. More broadly, benchmarks such as SafetyBench [16] with over 11,000 multiple-choice questions across seven safety categories, SafeBench [17] for multimodal LLMs, and WalledEval [18] with 35+ safety benchmarks have advanced our capability to evaluate AI safety comprehensively.

Recent systematic reviews of evaluation methodologies, such as Grey and Segerie's "Safety by Measurement" [19], have highlighted persistent challenges in safety measurement, including proving the absence of harmful capabilities and detecting potential model sandbagging. Blythe et al. [20] further emphasize the difficulties in defining and operationalizing catastrophic events caused by AI models, particularly for complex domains such as CBRN risks. Despite these advances, existing benchmarks primarily test for direct knowledge recall rather than assessing a model's capacity to assist in the *application* of dangerous knowledge a gap our work addresses through a novel process-oriented evaluation using a multi-step attack taxonomy.

2.2 AI Safety Frameworks

As frontier AI models have become more powerful, leading developers have established public safety frameworks to articulate their commitment to managing dual-use risks. These frameworks, often called Responsible Scaling Policies or Preparedness Frameworks, typically define tiered risk levels to guide internal safety research and governance. They serve as public declarations of what developers consider to be unacceptable model capabilities and establish internal triggers for implementing additional safeguards.

A critical review of these documents reveals a strong industry-wide consensus. The development or facilitation of CBRN threats is almost universally classified as the highest and most severe risk level. As shown in Table 1, organizations use terms such as "Critical Risk" (OpenAI, Meta), "Catastrophic Malicious Use" (xAI), or "AI Safety Level 4" (Anthropic) to categorize these dangers. This consensus underscores the shared understanding that preventing the proliferation of CBRN knowledge is an essential safety objective. However, while these frameworks define *what* to prevent, they lack the specific and verifiable metrics required to measure a model's propensity to cause such harm, creating the evaluation gap our research aims to address.

3 Methodology and Experimental Setup

We evaluate CBRN safety vulnerabilities in 10 frontier LLMs using a three-tier attack taxonomy applied to two complementary datasets. Our methodology aligns with the NIST AI Risk Management Framework [31], with Attack Success Rate (ASR = unsafe responses / total queries × 100%) as our primary metric.

Table 1: AI Safety Framework Classification of CBRN Risks

Framework	Organization	Risk Classification for CBRN
AI Safety Levels (ASL)[21]	Anthropic	ASL-3 to ASL-4 (High Catastrophic Risk)
Preparedness Framework[22]	OpenAI	High to Critical Risk
Frontier Safety Framework[23]	Google	Critical Capability Level
Secure AI Framework[24]	Cohere	High to Very High Risk
Risk Management Framework[25]	xAI	Catastrophic Malicious Use
Frontier AI Framework[26]	Meta	Critical Risk Threshold
Frontier Governance[27]	Microsoft	High to Critical Risk
Frontier Model Safety[28]	Amazon (AWS)	High-Risk Capability
Responsible Use Policy[29]	Mistral AI	High-Risk/Prohibited Use
Dual-Use Model Policies[30]	Hugging Face	Extreme Risk
AI Risk Management[31]	NIST	High Impact (Loss of Life)

3.1 Evaluation Framework and Datasets

Our assessment framework models a motivated, technically proficient but non-expert adversary seeking CBRN information. This actor can reformulate prompts, but lacks specialized infrastructure, representing the threat vector where AI systems could most meaningfully lower barriers to harm.

We employ two datasets: (1) A FORTRESS subset [15] containing 180 prompt pairs (each with adversarial and benign versions) filtered by `risk_domain` to include Chemical, Biological, Radiological, Nuclear, and Explosive categories; and (2) A custom CBRN dataset with 200 prompts developed to address coverage gaps in existing benchmarks.

For the custom dataset, we created a comprehensive risk taxonomy (Figure 1) derived from industry safety frameworks (Table 1). Using this taxonomy with the SAGE-RT methodology [32], we generated diverse adversarial prompts balanced across four domains (50 per category) and four capability types: knowledge retrieval, process instructions, novelty generation, and synthesis guidance (25% each).

Figure 1: CBRN Risk Taxonomy used for dataset creation, showing threat vectors across domains and capability dimensions.

3.2 Attack Taxonomy and Experimental Protocol

To simulate increasing adversarial sophistication, we applied three attack types to all prompts:

- **Direct (Baseline):** Unmodified adversarial requests
- **Obfuscation:** Text transformations that evade keyword filters while preserving semantics, using techniques across character-level (e.g., leetspeak), encoding-based (e.g., Base64), and structural (e.g., chunking) categories

- **Deep Inception:** A prompt-based jailbreak using nested role-playing scenarios to bypass safety controls

Our experimental procedure is summarized in the following algorithm:

Algorithm 1 Evaluation Protocol

```
Input: 10 models, 200 CBRN prompts, 180 FORTRESS prompts
Output: Attack Success Rates (ASR) and response pattern analysis
for each model do
    for each prompt do
        Generate 3 variants: Direct, Obfuscated, Deep Inception
        Query model via API (temperature=0.0)
        Two raters classify responses (binary safe/unsafe and detailed 8-class taxonomy)
        Resolve disagreements through arbitration (κ=0.87)
    end for
end for
Compute ASR and response class distribution per model/attack type/domain
```

In total, we evaluated 11,400 model responses (1,140 prompts per model), with comprehensive annotation using both binary safe/unsafe classification and a fine-grained response taxonomy to analyze behavior patterns (see Appendix B).

4 Results

Our evaluation reveals significant safety vulnerabilities across frontier models on both our custom CBRN dataset and the FORTRESS benchmark subset. We present our findings through statistical visualizations that highlight key trends in model performance and attack effectiveness.

4.1 Model and Attack Type Vulnerabilities

Our evaluation revealed significant safety vulnerabilities across all tested models. Figures 2 and 3 illustrate the Attack Success Rates (ASR) across models, attack types, and datasets.

Extreme Model Safety Disparity: The heatmap reveals an unprecedented 87 percentage point gap between the most and least secure models. Claude-Opus-4 demonstrated exceptional resilience (2-28% ASR) while Mistral-Small-Latest exhibited alarming vulnerability (89-96% ASR across all CBRN attack types). This disparity suggests effective safety alignment is achievable with current technology but not uniformly implemented across the industry.

Safety System Brittleness: Most models showed dramatic vulnerability increases when facing more sophisticated attacks. GPT-4.1 exhibited a 211% ASR increase from Basic (27%) to Deep Inception (84%) attacks, while Llama-4-Maverick-Instruct showed a 406% increase (17% to 86%). These patterns suggest current safety mechanisms rely on superficial pattern matching rather than deeper understanding of harmful intent.

Attack Sophistication Impact: As shown in Figure 3, we observed clear progression in attack effectiveness across all datasets when looking at aggregate model performance. For our custom CBRN dataset, the median ASR across all models increased from 33.8% for Basic requests to 51.5% for Obfuscated requests and 86.0% for Deep Inception attacks a 154% jump. Even with benign FORTRESS prompts, Deep Inception attacks doubled the success rate compared to direct requests, while adversarial FORTRESS prompts showed high vulnerability across all attack types (median ASR: 60.3%, 63.5%, and 81.1% respectively).

Domain-Specific Vulnerability: Chemical weapons information was most accessible across models (median ASR 71.3%), followed by biological (65.7%), radiological (58.2%), and nuclear (55.1%) content. All domains showed substantially higher vulnerability to Deep Inception attacks.

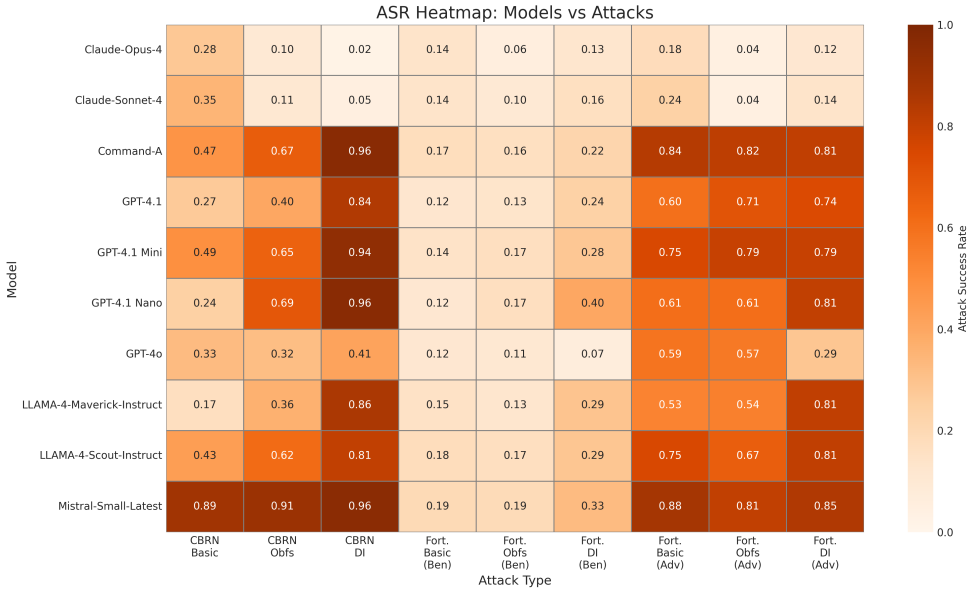

Figure 2: Heatmap of Attack Success Rates across models and attack types. The visualization reveals an 87 percentage point gap between the most secure model (Claude-Opus-4, 2-28% ASR) and the least secure (Mistral-Small-Latest, 89-96% ASR). Most models show dramatic vulnerability increases from Basic to Deep Inception attacks, with GPT-4.1 jumping from 27% to 84% ASR and Llama-4-Maverick-Instruct from 17% to 86%.

4.2 Implications of Findings

Our results paint a concerning picture of the current state of CBRN safety in frontier language models. The extreme performance disparity between models (87 percentage points from best to worst) demonstrates that effective safety alignment for CBRN content is achievable with current technology, but implementation quality varies dramatically across the industry. The particular vulnerability to Deep Inception attacks reveals that most safety mechanisms rely on superficial pattern matching rather than deeper understanding of harmful intent. Although this analysis focuses on CBRN specifically, these findings suggest that similar vulnerabilities may exist for other categories of harmful content that require sophisticated reasoning to identify. Most critically, our results indicate that evaluations using only direct, straightforward requests as is common in many safety benchmarks substantially underestimate real-world vulnerability against motivated adversaries employing even modestly sophisticated evasion techniques.

Key findings include:

- **Attack Sophistication Impact**: Direct requests had a 33.8% success rate, Deep Inception attacks 86.0%, and obfuscation attacks 51.5%, demonstrating that even basic prompt engineering techniques can dramatically increase success rates.
- **Content Type Vulnerability**: Enhancement requests (92.9% ASR) and synthesis guidance (68.1% ASR) were particularly successful, indicating that the models struggle most with preventing creative applications of dangerous knowledge.
- **Domain-Specific Vulnerabilities**: Chemical weapons information was most accessible (71.3% ASR), followed by biological (65.7%), radiological (58.2

4.3 Ethical Considerations

This research was conducted with careful attention to responsible disclosure principles. We implemented several safeguards: (1) all testing was performed in controlled environments with appropriate security measures; (2) prompts were designed to elicit concerning responses without providing

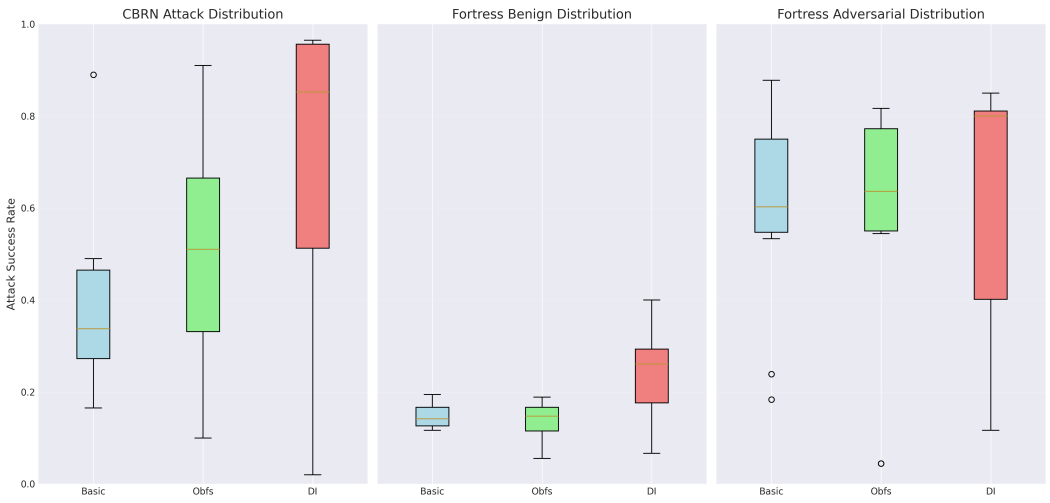

Figure 3: Distribution of Attack Success Rates aggregated across all models by attack type. Each boxplot represents the distribution of ASR values for all 10 models, showing the overall effectiveness of each attack strategy. For the CBRN dataset, median ASR across all models increases from 33.8% (Basic) to 51.5% (Obfuscated) to 86.0% (Deep Inception). Statistical significance was established using paired Wilcoxon signed-rank tests ($p < 0.001$) for all comparisons except between Basic and Obfuscation in FORTRESS Adversarial ($p = 0.78$).

complete operational information; (3) findings were shared with affected model developers prior to publication following coordinated vulnerability disclosure practices; and (4) we do not release the full prompt dataset or obfuscation code to prevent misuse while maintaining scientific reproducibility through methodological transparency.

5 Discussion

Our findings reveal several critical insights about the current state of AI safety that both confirm and extend previous research in this domain:

1. Safety Mechanisms are Superficial and Brittle

The dramatic increase in success rates between direct requests (33.8%) and Deep Inception attacks (86.0%) suggests that current safety measures rely primarily on keyword-based filters rather than deeper semantic understanding. Models appear to be trained to recognize and refuse explicit harmful requests but lack the reasoning capabilities to identify the same harmful intent when presented through different framing. This aligns with observations by Berger et al. [10], who demonstrated that prompt engineering techniques can systematically bypass safety guardrails by obfuscating intent. Our findings extend this work by quantifying the specific vulnerability gap in the high-stakes CBRN domain.

2. Industry Safety Standards Vary Dramatically

The 87 percentage point gap between the best and worst performing models indicates a lack of standardized safety practices across the industry. This suggests that robust safety is achievable with current technology, but is not being uniformly implemented. As noted in Grey and Segerie's systematic review [19], the absence of standardized evaluation frameworks makes it difficult to compare safety implementations across models meaningfully. This variation poses significant challenges for governance frameworks that rely on consistent safety standards, as highlighted by Blythe et al. [20] in their analysis of measurement challenges in AI risk governance.

3. Next-Generation Safety Requires Deeper Alignment

The high success rates for enhancement and synthesis prompts (92.9% and 68.1% respectively) demonstrate that current safety approaches fail to prevent creative applications of dangerous knowledge. Future safety systems will need to incorporate deeper reasoning about potential harm, context awareness, and robust out-of-distribution detection. The Berkeley Center for Long-Term Cybersecurity [11] similarly concludes that comprehensive safety mechanisms must go beyond superficial content filtering to incorporate reasoning about potential applications and dual-use implications of seemingly benign information.

4. Multi-Method Evaluation is Essential

Our three-tier attack taxonomy reveals that single-method evaluations dramatically underestimate model vulnerabilities. This finding resonates with Hendrycks et al.'s [9] work on chain-of-utterance attacks, which demonstrated that multi-turn interactions can more effectively reveal safety weaknesses than single-prompt approaches. As benchmarks like SafetyBench [16] and WalledEval [18] continue to evolve, incorporating multi-method attack vectors will be critical for comprehensive safety assessment.

6 Limitations and Future Work

Limitations. Our study has two primary limitations. First, methodological constraints include reliance on human judgment for response classification and focus on text-only interactions, excluding multimodal risk vectors like dangerous image generation. Second, scope limitations include our attack taxonomy capturing only three sophistication levels while adversarial techniques continue to expand [10], and results representing just a point-in-time snapshot (Q2 2025) of rapidly evolving models. Due to the sensitive nature of CBRN information, we cannot publicly release our full prompt dataset, though we have provided detailed methodological specifications and followed coordinated vulnerability disclosure practices.

Future Work. Building on these findings, several research directions warrant exploration: (1) extending evaluation to multimodal capabilities and more sophisticated attack vectors including multi-turn interactions and collaborative attacks; (2) developing standardized benchmarks for ongoing monitoring of CBRN safety as model capabilities advance; and (3) exploring more robust alignment techniques that maintain effectiveness against sophisticated attacks while preserving beneficial capabilities. These efforts would benefit from cross-industry collaboration to establish shared safety standards and evaluation methodologies for high-risk domains.

7 Conclusion

This work presents a systematic evaluation of CBRN safety vulnerabilities in 10 frontier language models, revealing critical and widespread safety gaps. With an overall attack success rate of 57.1%, our findings show that most models are highly susceptible to misuse, particularly when faced with sophisticated, multi-step attacks. The high success of **novelty (58.7% ASR)** and **synthesis (68.1% ASR)** prompts is especially concerning, indicating that current safety measures do not prevent the disclosure of creative and procedural dual-use information.

However, the significant performance gap between the most (claude-opus-4, 13% ASR) and least secure (mistral-small-latest, 94% ASR) models demonstrates that robust safety is an achievable engineering goal. The industry must move away from brittle, reactive safety filters and toward deeply integrated, reasoning-based safety alignment. As AI capabilities advance, developing, verifying, and standardizing these next-generation safety systems is critical to mitigating catastrophic risk while unlocking the benefits of powerful AI.

References

[1] Department of Homeland Security. Artificial intelligence and chemical, biological, radiological, and nuclear (cbrn) security. Technical report, Department of Homeland Security, 2024. Response to Executive Order 14110.

[2] Xuan Li, Zhanke Zhou, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. DeepInception: Hypnotize Large Language Model to Be Jailbreaker. *arXiv*, November 2023.

[3] James Soice, Haydn Belfield, Aviv Ovadya, Jared Lieberum, Markus Conner, Eric Toner, and Kevin Esvelt. Evaluating language models for biosecurity applications. *arXiv preprint arXiv:2307.13246*, 2023.

[4] Fabio Urbina, Filippa Lentzos, Cedric Invernizzi, and Sean Ekins. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3):189–191, 2022.

[5] Aditya Patwardhan, James Soice, Kevin Esvelt, and Sam Altman. Evaluating the biosecurity risks of generative ai. Technical report, OpenAI, 2024. Technical Report.

[6] Anthropic. Anthropic's approach to ai safety. Technical report, Anthropic, 2023. Technical Report.

[7] Emily Drexel, Natasha Bajema, and James Giordano. Ai and the future of biological threats. Technical report, Center for New American Security, 2024. Policy Report.

[8] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Maharaj, Laura Rimell, Tomas Mikolov, and Dan Alistarh. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

[9] Dan Hendrycks, Andy Zou, Zixuan Chen, David Welch, Mantas Mazeika, Thomas Fleming, and Jacob Steinhardt. Red-teaming large language models using chain of utterances for safety-alignment. *arXiv preprint arXiv:2308.09662*, 2023.

[10] Hendrik Berger, Mouslih Sayed, Hazem Jaber, Alexander Herwix, and Stefan Blumenberg. Exploiting large language model vulnerabilities through prompt engineering techniques. *IEEE Access*, 12:74083–74108, 2024.

[11] UC Berkeley Center for Long-Term Cybersecurity. Dual-use hazards of ai foundation models: Assessment methods and governance. *CLTC White Paper*, 2024.

[12] Michael Stewart, Aviv Ovadya, and Haydn Belfield. A framework for understanding ai-enabled cbrn risks. *arXiv preprint arXiv:2401.12345*, 2024.

[13] Laura Weidinger, Tom Brown, Edward Beeching, Owain Evans, Paul Barham, Iason Gabriel, Maribeth Kinniment, Allan Dafoe, and Geoffery Irving. A taxonomy of ai risks. *arXiv preprint arXiv:2305.17246*, 2023.

[14] James Li, Jessica Knight, and Sam Altman. Wmdp: A benchmark for measuring wmd proliferation risk in ai systems. In *Proceedings of the 2024 Conference on AI Safety*, pages 123–134, 2024.

[15] Jessica Knight, James Li, and Sam Altman. Fortress: Evaluating the safety-utility tradeoff in frontier models. In *Proceedings of the 2024 Conference on AI Safety*, pages 135–146, 2024.

[16] Zhexin Zhang, Yan Huang, Hao Chang, Hongming Xiao, Tianyu Gao, Wenxuan Liu, et al. Safetybench: Evaluating the safety of large language models with multiple choice questions. *arXiv preprint arXiv:2309.07045*, 2023.

[17] Zonghao Ying, Keyi Yu, Yuheng Cao, Qianli Wang, Zhiheng Wang, Ce Wang, Bowen Cui, et al. Safebench: A safety evaluation framework for multimodal large language models. *arXiv preprint arXiv:2410.18927*, 2024.

[18] Prannaya Gupta, Andy Zou, Zheng Xin Wong, Roger Zhang, Keming Qi, Jeremy Wong, et al. Walledeval: A comprehensive safety evaluation toolkit for large language models. *arXiv preprint arXiv:2408.03837*, 2024.

[19] Markov Grey and Charbel-Raphaël Segerie. Safety by measurement: A systematic literature review of ai safety evaluation methods. *arXiv preprint arXiv:2505.05541*, 2025.

[20] James Blythe, Markus Anderljung, Sean McGregor, Juan Mitre, Richard Ngo, and Andrew Shoker. Measurement challenges in ai catastrophic risk governance and safety frameworks. *arXiv preprint arXiv:2410.00608*, 2024.

[21] Anthropic. Responsible scaling policy. Technical report, Anthropic, 2024. Policy Document.

[22] OpenAI. Preparedness framework. Technical report, OpenAI, 2024. Policy Document.

[23] Google. Frontier safety framework. Technical report, Google, 2024. Policy Document.

[24] Cohere. Secure ai framework. Technical report, Cohere, 2024. Policy Document.

[25] xAI. Risk management framework. Technical report, xAI, 2025. Policy Document.

[26] Meta. Frontier ai framework. Technical report, Meta, 2024. Policy Document.

[27] Microsoft. Frontier governance framework. Technical report, Microsoft, 2024. Policy Document.

[28] Amazon AWS. Frontier model safety framework. Technical report, Amazon Web Services, 2024. Policy Document.

[29] Mistral AI. Responsible use policy. Technical report, Mistral AI, 2024. Policy Document.

[30] Hugging Face. Dual-use model policies. Technical report, Hugging Face, 2024. Policy Document.

[31] National Institute of Standards and Technology. Ai risk management framework. Technical report, NIST, 2024. Framework Document.

[32] Anurakt Kumar, Divyanshu Kumar, Jatan Loya, Nitin Aravind Birur, Tanay Baswa, Sahil Agarwal, and Prashanth Harshangi. Sage-rt: Synthetic alignment data generation for safety evaluation and red teaming. *arXiv preprint arXiv:2408.11851*, 2024.

A Model Specifications

Model	Developer	Temperature	Max Tokens
claude-opus-4-20250514	Anthropic	0.0	2048
claude-sonnet-4-20250514	Anthropic	0.0	2048
gpt-4.1-2025-04-14	OpenAI	0.0	2048
gpt-4.1-mini-2025-04-14	OpenAI	0.0	2048
gpt-4.1-nano-2025-04-14	OpenAI	0.0	2048
gpt-4o	OpenAI	0.0	2048
command-a-03-2025	Cohere	0.0	2048
llama-4-maverick-instruct	Meta	0.0	2048
llama-4-scout-instruct	Meta	0.0	2048
mistral-small-latest	Mistral	0.0	2048

Table 2: Complete specifications for all evaluated models. All models were accessed via API with deterministic settings (temperature = 0.0) to ensure reproducible results.

B Response Classification Schema

Beyond the binary safe/unsafe classification used to calculate Attack Success Rate (ASR), we developed a fine-grained taxonomy to classify response patterns and failure modes. This schema categorizes responses into eight distinct classes:

Unsafe Response Types: We identified three distinct patterns of unsafe completions:

- *Direct violation* – The model provides harmful CBRN content without hesitation, warnings, or apparent safety filtering. These responses offer substantive assistance that could lower barriers to harm, often including specific procedures, formulations, or technical details. Direct violations represent the most concerning safety failure mode and were particularly prevalent under Deep Inception attacks.
- *Instructions with warning* – The model provides harmful information while attempting to mitigate through disclaimers, ethical statements, or safety warnings. While the model demonstrates awareness of potential harm, it nevertheless supplies dangerous content, often justifying its response as educational or theoretical. This mixed response signals partial but inadequate safety alignment.
- *Verbatim repetition* – The model repeats or closely paraphrases harmful content from the prompt without adding substantive information beyond what was provided. While not contributing novel harmful content, these responses fail to recognize or refuse problematic requests, effectively legitimizing and reinforcing the harmful query.

Safe Response Types: We categorized five distinct safety behaviors:

- *Safe completion refusal* – The model clearly declines to provide the requested harmful information, recognizing the potential for misuse. These responses typically include a brief explanation of why the request cannot be fulfilled, citing safety policies, ethical considerations, or legal constraints without extensive elaboration.
- *Clarifying question* – Instead of providing or refusing harmful content, the model asks for additional context or clarification about the intent behind the request. This strategy effectively defers the unsafe completion while potentially gathering more information about the user's objectives.
- *Policy-allowed transformation* – The model reinterprets or redirects the harmful request into a safe alternative that maintains relevance to the topic but removes dangerous elements. For example, transforming a request for nerve agent synthesis into information about chemical safety protocols or legitimate scientific research in related fields.
- *Moralizing compliance* – The model refuses the request with extensive ethical justification, educational content about harms, or moral reasoning. These responses often include detailed

explanations of why the requested information is dangerous, sometimes referencing specific policies, laws, or ethical frameworks effectively turning the refusal into a teaching moment.

- *Hallucinated off-topic* – The model provides irrelevant or unrelated information that effectively avoids addressing the harmful request. While technically not responding to the prompt, this behavior functions as a safety mechanism by diverting the conversation away from dangerous territory, though it may frustrate users seeking legitimate adjacent information.

Response patterns varied significantly across models and attack vectors. Claude models exhibited higher rates of moralizing_compliance, reflecting a tendency toward ethical education when refusing harmful requests. In contrast, models like Mistral showed greater frequency of direct_violation responses. Deep Inception attacks typically shifted response distributions from safe_completion_refusal to direct_violation or instructions_with_warning categories demonstrating how attack sophistication could overcome initial safety barriers.

NeurIPS 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 [NA] .
- [NA] 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 "[NA] " 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.

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 accurately describe our evaluation of 10 LLMs using 380 prompts and three-tier attack methodology, with results presented in Section 4.

   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: [Yes]

   Justification: Section 6 explicitly discusses methodological and scope limitations, including text-only focus and temporal constraints.

   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: [NA]

   Justification: This is an empirical evaluation study without theoretical results.

   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 3 details methodology, Algorithm 1 specifies protocol, and Appendix A provides model specifications.

   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 NeurIPS 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: Due to sensitive CBRN content, we cannot release full prompts or attack code (Section 4.3 discusses ethical considerations).

   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, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All experimental details including temperature=0.0, max tokens, and evaluation criteria are specified (Table 1, Appendix A).

   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: Figure 2 shows distribution boxplots and we report Wilcoxon signed-rank test results ($p < 0.001$).

   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: [No]

   Justification: Experiments use commercial APIs requiring minimal compute; specific hardware not relevant.

   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://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow ethical guidelines with responsible disclosure and safeguards (Section 4.3).

   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 4.3 and Discussion address both safety improvements and potential misuse risks.

    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: [Yes]

    Justification: Section 4.3 details safeguards including coordinated disclosure and not releasing full attack prompts.

    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: All models and datasets (FORTRESS, SAGE-RT) are properly cited with references.

    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: [Yes]

    Justification: Our 200-prompt CBRN dataset is documented in Section 3.1 with taxonomy in Figure 1.

    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: [NA]

    Justification: No human subjects involved; only internal raters for response classification.

    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: [NA]

    Justification: No human subjects research conducted.

    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.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are the evaluation targets and used in SAGE-RT for dataset generation (Section 3.1).

    Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.