QUARL: Quantifying Adversarial Risks in Language Models

Joshua Ackerman[1], **George Cybenko**[1], **Paul Lintilhac**[1], **Henry Scheible**[1], **Nathaniel D. Bastian**[2]

[1] Thayer School of Engineering and Department of Computer Science, Dartmouth College, Hanover, NH 03755 USA
[2]Department of Electrical Engineering and Computer Science, United States Military Academy, West Point, NY 10996 USA
joshua.m.ackerman.gr@dartmouth.edu, george.cybenko@dartmouth.edu, paul.s.lintilhac.th@dartmouth.edu,
henry.j.scheible.26@dartmouth.edu, nathaniel.bastian@westpoint.edu

Abstract

It is well documented that artificial intelligence (AI) systems have various types of vulnerabilities and associated risks. As such systems are deployed in safety-critical domains, it has become necessary not only to identify and enumerate the vulnerabilities but also to quantify the resulting risks. In this position paper, we discuss approaches for the challenge of quantifying AI risks. The approach is based on a general framework for testing and evaluating language model systems that we have previously developed (called TEL'M). In particular, we extend TEL'M to deal with the problem of quantifying the effort required by an adversary to discover and exploit a language model vulnerability.

Introduction

The U.S. National Institute of Standards and Technology has released the *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST 2024). This is a comprehensive proposal that outlines an artificial intelligence (AI) risk management methodology and associated processes for executing the methodology. It should be noted that there are other AI risk management proposals that have been developed by the UK and EU for example (Neuman et al. 2021).

However, as with many risk management activities, such as those that occur in finance, transportation and national security for example, the actual quantification of risk is not prescribed even though quantification is what makes risk assessment actionable. The quantification details are typically "left to the reader."

According to standard terminology, the "risk" of an event occurring is the product of the probability that the event does happen times the cost of the event's consequence. This is articulated in the AI RMF as:

> "In the context of the AI RMF, risk refers to the composite measure of an event's probability of occurring and the magnitude or degree of the consequences of the corresponding event. The impacts, or consequences, of AI systems can be positive, negative, or both and can result in opportunities or threats (Adapted from: ISO 31000:2018). When considering the negative impact of a potential event, risk is a function of 1) the negative impact, or magnitude of harm, that would arise if the circumstance or event occurs and 2) the likelihood of occurrence."

In this approach, the "event's probability" and "the magnitude or degree of the consequences of the corresponding event" are typically taken from the asset owners' perspectives. For example, an adversary might be able to jailbreak a language model to produce responses that violate safety or bias constraints (Robey et al. 2023). The owner/operator of the language model would then suffer various consequences such as legal liabilities, regulatory violations and/or negative brand impacts. Such consequences can be quantified numerically using classical utility theory (Fishburn 1968). However, the owner/operator perspective is silent about the costs of creating attacks against a language model from the adversaries' points of view.

In this paper, we introduce QUARL (Quantifying Adversarial Risks in Language Models) as an approach to quantifying *attacker* costs for creating an attack. Specifically, we consider adversarially induced events as opposed to benignly occurring events such as a language model system making an error as a result of standard use. Our goal is to inform the owner/operator of a language model system how possible attacks could be cost-effectively carried out by different classes of adversaries (such as nation states or criminal cartels) (Sigholm 2013; Yeo, Birch, and Bengtsson 2019).

The QUARL Approach

There is a large space of adversarial attacks that can culminate in the exploitation of a language model system. Commonly discussed examples of such attacks include, but are not limited to:

- **Membership Inference:** An adversary infers the inclusion of a sensitive data point in the model (Shokri et al. 2017);
- **Data Reconstruction:** An adversary reconstructs a sensitive data point (Guo et al. 2022);
- **Input Attack:** An adversary tailors an input or physically alters the environment to subvert or reprogram the system (Xie et al. 2019);

- **Model Theft:** An adversary learns proprietary hyperparameters or model architecture choices (Juuti et al. 2019);
- **Exploitable Adversarial Examples:** An adversary discovers an input which produces an incorrect response that the adversary can exploit (Yuan et al. 2019).

Other attacks on machine learning systems are known and are still being discovered (Kurakin, Goodfellow, and Bengio 2016; Huang et al. 2011; Hu et al. 2022; Schwinn et al. 2023; Robey et al. 2023). For sure, the impact of an adversarial attack occurring is application and stakeholder specific. However, we argue that the tactics, techniques, and procedures (TTPs) to model the risks of any of these attacks occurring against a specific model are sufficiently generic and can be performed in a rigorous fashion using QUARL.

The QUARL framework builds on the TEL'M (Test and Evaluation of Language Models) approach (Cybenko, Ackerman, and Lintilhac 2024). TEL'M was originally designed for testing and evaluating quantitative metrics for "friendly" use-case properties of language model systems. We consider friendly properties to be properties that are desired by the developer/operator of a language model - namely properties like accuracy, fairness, robustness, etc. The key steps of TEL'M are:

TEL'M Steps

1. Step #1: Identify the model's underlying *task*.
2. Step #2: Identify the *properties* of interest;
3. Step #3: Identify the *metrics* to quantify the properties;
4. Step #4: Do the above in the context of principled *experimental design*;
5. Step #5: *Perform and document* the experiments.

Details fleshing out the above steps can be found in the foundational TEL'M document (Cybenko, Ackerman, and Lintilhac 2024).

QUARL has the same framework as TEL'M but views properties and metrics from the attacker's point of view. That is, a possible QUARL property could be "The existence of a specific Membership Inference Attack against the model." Recall that a Membership Inference Attack seeks to determine, with high confidence, whether a specific data item or set of data items was included in the training or fine-tuning of a language model.

Another fundamental difference between TEL'M and QUARL is that the metrics of an attack (property) quantify the effort and access required to conduct the attack.

QUARL Steps (differences with TEL'M in bold)

1. Step #1: Identify the model's underlying *task*.
2. **Step #2**: Identify the ***attack*** of interest;
3. **Step #3**: Identify the ***metrics*** to quantify the ***creation and execution of an attack***;
4. Step #4: Do the above in the context of principled *experimental design*;
5. Step #5: *Perform and document* the experiments.

The attacks we consider here are attacks on the training data, input-output semantics and possible guardrails of the language model tasks, not generic attacks against the underlying software systems on which a language model operates. For example, we do not consider buffer overflow attacks against a web browser implementing a language model's interface with remote users in QUARL, deferring those types of attacks to the classical cybersecurity literature.

On the other hand, crafting prompts to a language model that result in "toxic" responses (Deshpande et al. 2023) for which guardrails were put in place by the owner/operator of a language model to avoid qualifies as an attack of interest in the QUARL context.

The concept of a "Cyber Kill Chain" was developed originally by Lockheed-Martin and subsequently refined by MITRE to model the steps required to conduct a classical cyber attack against a computer system (CYBOTS 2018). The Cyber Kill Chain involves multiple steps, conducted sequentially, starting with reconnaissance of the target system and ending with actions on objectives. The difficulty of each step can be quantified in terms of the estimated attacker effort, capabilities and cost needed to execute the whole chain.

By contrast, QUARL proposes metrics related to the various accesses, compute power, and time required to conduct an adversarial attack against a language model in the sense described here. Example QUARL metrics include, but are not limited to:

- **Required access to training data** - Some attacks, particularly membership inference attacks, require knowledge of the precise training data size or knowledge of the distribution from which the training data on which the language model was trained. For example, creating a surrogate training data based on ChatGPT's crawling of text from the World Wide Web may be prohibitively expensive for most attackers yet feasible for nation states or large commercial concerns. On the other hand, knowing that a model was trained on a CIFAR dataset (Krizhevsky, Nair, and Hinton 2009) could make access to the training data scalable and simple. Access to and knowledge of any fine-tuning datasets is a similar consideration which we do not break out separately here;
- **Required access to trained model** - In addition to or instead of access to training data, an adversarial attack against a language model may require some number of prompt-response interactions with that model. For example, crafting prompts to jailbreak guardrails that are aimed to prevent toxic or otherwise inappropriate responses may require a large number of tries before succeeding (Robey et al. 2023). If such accesses are against a commercial system charging per access or a closed, propriety system such as military or company-internal that might monitor usage closely, some cost can be attached, whether in financial terms or discovery probabilities;
- **Required compute power** - Current techniques for membership inference are basically "proof-of-concept" demonstrations based on mathematical arguments (Ilyas et al. 2022; Carlini et al. 2022). Some approaches require

training N models where N is the size of the training data. If training a single model requires, say a week of compute time on reasonable available infrastructure, and $N \approx 10^6$, the compute power is far beyond what is available to attackers not aligned with a nation state or a well-funded commercial concern. The cost and accessibility or required compute power is time-varying, typically decreasing so it should be expressed in compute, memory and time units such as priced by Google's COLAB infrastructure (Goggle 2024);

- **Adversary capabilities** - In general, published attack techniques tend to be highly technical. While there are often open source implementations of some attack techniques, even those techniques require some programming skill to port and implement at scale. Moreover, nation state and well-funded non-nation state actors can employ scientists who can discover and keep confidential attack techniques that are far more reliable, scalable and cost-effective that published attacks again language models. The possible existence of "zero day" attacks suggest that any quantitative metric estimates based on published or red team implementations of attacks just serve as upper bounds to the cost, efforts and accesses required in a QUARL risk quantification exercise. Related efforts to quantifying attacker capabilities can be leveraged (ben Othmane et al. 2015);
- **Exploitability** - By definition, the documented cases of adversarial attacks against language models show that with appropriate knowledge, access and compute power, the attacks can be successful with probability one. However, the key question for a potential attacker is whether an attack can actually be exploitable. For example, demonstrating that by changing only a few humanly-imperceptible pixels in an image of a panda, an image classifier will classify the image as a baboon is interesting. But is it useful to an attacker trying to exploit a specific image for a specific desired outcome with some robustness and noise tolerance?

These are only some obvious and easily quantifiable metrics that can be used by QUARL. In the case of each metric, the quantification must be understood to be in terms of stochasticity and random variables. That is to say, the computing effort required may be a random variable depending on luck in training and approach. As such, some proposed metrics may require multiple samples to achieve robust estimates of the metric. More comprehensive discussions of various types of possible metrics (simple, compound and higher-order) that are possible within QUARL as well as sampling theory required to achieve desired confidences are described in the TEL'M documentation (Cybenko, Ackerman, and Lintilhac 2024) and are outside the scope of this present contribution.

Future Considerations

Previous approaches for quantifying work effort and time in cybersecurity settings, for example software protection, may be applied in the QUARL context as well (Carin, Cybenko, and Hughes 2008). In particular, it could be possible to use so-called information markets to elicit various estimates of metrics with incurring the possible enormous costs of computing many trial runs of training. In fact, information markets are often used to estimate various otherwise difficult to measure quantities in industry (Hanson 2003).

Another important line of investigation involves the formulation of adversarial attacks against language models in game theoretic terms. Briefly, in a security risk assessment problem in which the attack surface is large and complex, the defender has to balance investments in defenses for the possible attacks and their consequences. At the same time, the adversary has to decide which attack achieves some combination of desired end effects, all within budgetary constraints on both sides. This has been explored in traditional cybersecurity contexts (Cybenko et al. 2019).

A formulation in terms of game theory and decision theory, such as adversarial risk analysis (ARA) (Banks et al. 2020), that builds on QUARL would be powerful and useful to concerned parties, we believe.

Moreover, ARA is a methodology for quantifying utilities, which is a major challenge in classical risk and game theory approaches, as is the issue of what constitutes common knowledge among the players. Successful applications of ARA to security applications have been demonstrated (**?**) and we believe QUARL can be used in similar ways.

Summary

QUARL builds on the same disciplined process of TEL'M, but differs in focus and motivation. The staging of TEL'M is centered around the test and evaluation of a language model for deliberate and aspirational properties such as accuracy during normal performance. In contrast, in QUARL we recenter the TEL'M framework around measuring the preservation of these user intended properties in an operational environment under adversarial attack.

A critical difference, between these perspectives is that the probability of the user intended properties withstanding this environment is not just the probability of an attack succeeding, but also the probability an adversary decides it is worthwhile to launch an attack. Accordingly, while the traditional measure of how successful these attacks are against a model remains important, here we advocate for further considering realistic metrics for the costs of these attacks, such as the access requirements, compute costs, and capabilities that an adversary requires to launch them.

We believe QUARL is a first step towards quantifying adversarial risks for language models of various types, including multimodal language models and other AI embodiments.

Acknowledgments

This work was supported in part by the Defense Advanced Research Projects Agency under Cooperative Agreement No. HR00112420351, the U.S. Army Combat Capabilities Development Command Army Research Laboratory under Support Agreement No. USMA21050, and the U.S. Air Force Research Laboratory Autonomous Capabilities Team 3. The views expressed in this paper are those of the authors and do not reflect the official policy or position of the U.S. Department of Defense or the U.S. Government.

References

Banks, D.; Gallego, V.; Naveiro, R.; and Insua, D. R. 2020. Adversarial Risk Analysis (Overview). arXiv:2007.02613.

ben Othmane, L.; Ranchal, R.; Fernando, R.; Bhargava, B.; and Bodden, E. 2015. Incorporating attacker capabilities in risk estimation and mitigation. *Computers & Security*, 51: 41–61.

Carin, L.; Cybenko, G.; and Hughes, J. 2008. Cybersecurity strategies: The queries methodology. *Computer*, 41(8): 20–26.

Carlini, N.; Chien, S.; Nasr, M.; Song, S.; Terzis, A.; and Tramer, F. 2022. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, 1897–1914. IEEE.

Cybenko, G.; Ackerman, J.; and Lintilhac, P. 2024. TEL'M: Test and Evaluation of Language Models. *arXiv preprint arXiv:2404.10200*.

Cybenko, G.; Wellman, M.; Liu, P.; and Zhu, M. 2019. Overview of control and game theory in adaptive cyber defenses. *Adversarial and Uncertain Reasoning for Adaptive Cyber Defense: Control-and Game-Theoretic Approaches to Cyber Security*, 1–11.

CYBOTS. 2018. AN INTRODUCTION TO MITRE ATT&CK. https://cybotsai.com/introduction-mitre-attck/.

Deshpande, A.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; and Narasimhan, K. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

Fishburn, P. C. 1968. Utility theory. *Management science*, 14(5): 335–378.

Goggle. 2024. COLAB Pricing. ttps://colab.researc.google.com/signup.

Guo, C.; Karrer, B.; Chaudhuri, K.; and van der Maaten, L. 2022. Bounding training data reconstruction in private (deep) learning. In *International Conference on Machine Learning*, 8056–8071. PMLR.

Hanson, R. 2003. Combinatorial information market design. *Information Systems Frontiers*, 5: 107–119.

Hu, H.; Salcic, Z.; Sun, L.; Dobbie, G.; Yu, P. S.; and Zhang, X. 2022. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s): 1–37.

Huang, L.; Joseph, A. D.; Nelson, B.; Rubinstein, B. I.; and Tygar, J. D. 2011. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, 43–58.

Ilyas, A.; Park, S. M.; Engstrom, L.; Leclerc, G.; and Madry, A. 2022. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*.

Juuti, M.; Szyller, S.; Marchal, S.; and Asokan, N. 2019. PRADA: protecting against DNN model stealing attacks. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, 512–527. IEEE.

Krizhevsky, A.; Nair, V.; and Hinton, G. 2009. The CIFAR-10 dataset. https://www.cs.toronto.edu/~kriz/cifar.html.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

Neuman, K. L.; Cory-Wright, D.; Hespeler, C. B.; and White, M. 2021. European Commission's Proposed Regulation on Artificial Intelligence: Conducting a Conformity Assessment for High-Risk AI-Say What? *The Journal of Robotics, Artificial Intelligence & Law*, 5.

NIST. 2024. Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pd.

Robey, A.; Wong, E.; Hassani, H.; and Pappas, G. J. 2023. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*.

Schwinn, L.; Dobre, D.; Günnemann, S.; and Gidel, G. 2023. Adversarial attacks and defenses in large language models: Old and new threats. In *Proceedings on*, 103–117. PMLR.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, 3–18. IEEE.

Sigholm, J. 2013. Non-state actors in cyberspace operations. *Journal of Military Studies*, 4(1): 1–37.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2730–2739.

Yeo, S.; Birch, A. S.; and Bengtsson, H. I. J. 2019. The Role of State Actors in Cybersecurity: Can State Actors Find Their Role in Cyberspace? In *National Security: Breakthroughs in Research and Practice*, 16–43. IGI Global.

Yuan, X.; He, P.; Zhu, Q.; and Li, X. 2019. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9): 2805–2824.