A Framework to Evaluate the Risks of LLMs for Assisting CBRN Production Processes

CNS
Washington, DC Office

NONPRO NOTES
FEBRUARY 2024

Ian Stewart

Middlebury Institute *of*
International Studies at Monterey
James Martin Center for Nonproliferation Studies

James Martin Center for Nonproliferation Studies (Washington, DC Office)
Middlebury Institute for International Studies at Monterey

1400 K Street, NW, Suite 1225, Washington, DC 20005
Phone: +1 (202) 842-3100
www.nonproliferation.org/dc

The author is grateful to Natasha Bajema who is working on related topics at CNS for suggesting edits to a draft version of this report.

The views, judgments, and conclusions in this report are the sole representations of the author and do not necessarily represent either the official position or policy or bear the endorsement of CNS or the Middlebury Institute of International Studies at Monterey.

The cover image was generated by Dalle 3 based on a prompt to produce an image related to LLMs and CBRN proliferation. It purports to show a control room monitoring CBRN threats. It is notable that the text in the image is gibberish, as discussed in the paper.

Executive Summary

This paper examines how Large Language Models (LLMs) could contribute to the proliferation of chemical, biological, radiological, and nuclear (CBRN) weapons. A framework for examining the potential contribution of LLMs is presented. This framework identifies five areas of possible contribution, including brainstorming production processes, providing technical assistance, generating scripts or code for process simulation, aiding in designing relevant parts, and potentially linking to manufacturing services. It also outlines mitigation measures, distinguishing between built-in limitations of current LLMs and proactive strategies to prevent misuse. Finally, the paper underscores the importance of balancing the restriction of LLMs in CBRN applications with the preservation of their beneficial uses in advancing scientific research and development.

Contents

Note: This paper is part of broader work at CNS to understand the implications of generative design for proliferation challenges. Staff across CNS are undertaking work to understand the nexus between LLMs and proliferation and nonproliferation use cases. CNS is looking to engage with companies and regulators working in this space.

Introduction

Large Language Models (LLMs) or foundation models, such as GPT 4, have captured widespread attention primarily due to their ability to generate text about a seemingly unlimited range of topics. The impressive capability and anticipated trajectory of such models has led to concerns among tech developers and policymakers that this ability could be misused for negative purposes in several domains ranging from disinformation to the proliferation of chemical, biological, radiological, and nuclear (CBRN) weapons. While significant work is now underway to understand risk pathways and to develop evaluations to determine whether models are capable of problematic acts, so far the work in the CBRN domain has been relatively nascent, and much of the current work focuses specifically on chemical and biological areas. This paper proposes a framework for evaluating the risks of LLMs for CBRN proliferation more broadly by considering the different pathways for how LLMs could aid CBRN production processes.

This paper is based on the premise that consumer-facing Large Language Models (LLMs) should not contribute to the proliferation of Chemical, Biological, Radiological, and Nuclear (CBRN) weapons or to weapons proliferation more broadly. The production processes for CBRN and weapons are diverse. For nuclear weapons, for example, this includes mining uranium, milling, uranium enrichment, fuel fabrication, reactor operation, reprocessing spent fuel, designing and experimenting with weapons and explosives, and manufacturing the weapons themselves. Each of these steps is unique and requires a significant amount of scientific and engineering know-how. Many of these steps also benefit from modeling and simulation. While governments and defense companies may explore training niche LLMs to assist with CBRN and weapons production, private sector companies that offer LLMs as a for-profit service to the public and the business sector should take reasonable steps to ensure their models are not used for these purposes.

To this end, this paper outlines a framework to evaluate the risks that LLMs could contribute to CBRN proliferation. It then proposes mitigation measures to manage these risks, where they exist. The paper also highlights the potential downsides of constraining LLMs output related to CBRN topics; such actions will likely result in constraining legitimate uses of LLMs and limit their positive impact on advancing relevant scientific disciplines. Thus, careful thought is needed to ensure controls and restrictions are proportionate and effective.

CBRN Evaluation Framework

Given the technical differences across the production processes of CBRN, it is challenging to develop an inclusive framework that is applicable across nuclear, missile, chemical and biological proliferation issues and identifies the full range of use-cases of LLMs (or pathways as they are described below). For this, a nuanced technology-by-technology examination would be required. Moreover, a discussion about the risks of LLMs should also not be limited to production processes but ideally also account for the misuse of dual-use materials and technologies. For example, LLMs should not help individuals with gain-of-function research related to a viral pathogen.

Thus, in evaluating the risks of LLMs for assisting CBRN production processes and identifying the pathways below, four key questions should be kept in mind. The first is whether the LLM can produce any CBRN relevant output at all. The types of relevant output are examined further below. Second, is there a distinct and new role for the LLM, or is the LLM simply repackaging information that could be found online? A specific question here is whether the LLM holds and can impart tacit knowledge, which is often an essential barrier to a novice undertaking a complex task. The third is whether it would be a reasonable or proportionate expectation to prevent the LLM from producing that output. Fourth, are there legal restrictions that would prohibit the LLM from producing that output in some or all circumstances (e.g., such as export control considerations)? And finally, how practical is it to control the output, particularly given the increasing prevalence of open-source models?

Pathways

This section is a technology-agnostic effort to map out pathways through which LLMs can assist CBRN production processes. Future work should examine each weapon type separately with the purpose of identifying technology-specific pathways.

The first pathway examines whether the LLM can help to brainstorm a production process including breaking the process down into steps, identifying required materials, equipment, and expertise. This information is probably available in books and on the web. As such, while LLMs could provide assistance, the value-added by the LLM would be limited and likely nonconsequential. For this reason, it would be disproportionate to restrict LLMs from producing this type of output in most cases.

The second pathway is whether the LLM can provide technical assistance in the design, development, or manufacturing of a relevant item where the user could describe problems with implementation and upload images, videos, or log files for the LLM to diagnose. LLMs are presently capable of diagnosing text-based log information, but the imminent rise of multimodal models may mean that image- and video-based diagnostics may soon become feasible.

Technical assistance implies that the user can ask questions of the LLM to help overcome specific process challenges. For example, if an individual was machining a part from metal using a CNC router, technical assistance could involve questions about flaws in the finish of a machined part. A notable point here is that the provision of technical assistance for CBRN, weapons-related and dual-use technologies is export controlled. For this reason, providing technical assistance to foreign nationals through an LLM could be found to be a breach of export controls.

The third pathway is whether the LLM could produce scripts or code to evaluate designs or simulate process steps. This is likely to be one of the key areas where LLMs could provide assistance in CBRN production processes where computer design and simulation are essential steps. For example, hypersonic missile development relies on computational fluid dynamics and other forms of computer-based modelling; nuclear reactor design is reliant on monte carlo neutron flux modelling (among others); and missile flight modelling is an important step in missile development and deployment. Presently, engineers often consult specialist software and libraries for each of these applications, and some of these are open source while others are proprietary. It is conceivable that LLMs could generate scripts and code to enable this modelling and computer simulation. In the short term, it is perhaps more likely the case that LLMs will act as interfaces to existing software rather than directly replace the modelling capability of such software, but in the longer term LLMs may eliminate the need for specialist software

The fourth pathway is whether the LLM can aid in designing relevant parts. As of now, LLMs appear incapable of generating functional engineering designs. GPT 4 can integrate with services such as Dalle 3, which is an image generation tool, and Sora, which is a video generation tool. However, these tools appear to struggle with production of real-world features as demonstrated with the cover image for this paper, which was generated with Dalle 3 and contains gibberish text along with the images. Engineering design is fundamentally different from the production of artwork as the design must account for physics and engineering considerations of the real-world application. Despite the limitations of the current technology, it is possible that LLMs could generate a functional design for an aeroengine in the future. Should such capabilities emerge, an urgent priority should be ensuring that CBRN and weapons-related components cannot be designed by the LLM in the first place.

The fifth pathway is whether the LLM can send components for manufacture. Presently, LLMs cannot design real-world items let alone manufacture parts. If in the future LLMs can produce engineering designs, it seems inevitable that LLMs may be linked to manufacturing services. If and when this occurs, LLMs could in principle design a part and then have it manufactured. This raises obvious concerns for CBRN domain as the production of products, parts, or materials would directly increase the risk of CBRN proliferation. Ensuring that mitigation measures are in place to prevent production of CBRN parts should be a key priority.

Mitigation Measures

From this examination, there may be several areas where an LLM could aid CBRN production processes and exacerbate the risk of weapons proliferation—especially as the models improve rapidly and evolve to address more functional problems. This section examines approaches to mitigate these risks, which are categorized as built-in and proactive.

Built-in Mitigation Measures

The risks of LLMs for CBRN are already mitigated by several factors internal to the models themselves. The first and most obvious is not CBRN specific and relates to the current limitations of current LLMs (e.g., hallucinations). It would be foolhardy to rely on an LLM today for any safety or mission critical task.

Beyond this general point, other factors include the following. Firstly, LLMs have not been trained with the goal of creating computer models, simulations, and functional designs in mind. This may change in the future as companies building LLMs or the open-source community explore more niche markets and use-cases for the technology. Second, LLMs generally lack domain-specific insights; they can answer questions based on their training data. In the future, this may change either as models are fine-tuned on CBRN relevant data or as LLMs draw on domain specific data in responding to user prompts (i.e. through retrieval augmented generation). Third, even if LLMs can produce CBRN relevant outputs, it remains unclear whether these outputs would extend beyond what is already available online from reading the original data sources. This gets to the question of whether LLMs can hold and impart tacit knowledge—an area where much work is needed. Across these areas, it will be important to develop and maintain robust evaluations to provide the capabilities of LLM models for exacerbating CBRN proliferation. The major challenge with these built-in mitigations is that they may well disappear over time (even rapidly) as models improve and as the training domains for such models expand.

Proactive Mitigation Measures

While LLMs have current limitations on their ability to generate CBRN-relevant output, as outlined above, proactive mitigations are also necessary when the LLM can produce problematic output as the capability of LLMs improves. There are a number of proactive mitigation measures that can be pursued.

The first of these is the use of a machine learning classifier which sits on top of the LLM. This appears to be the approach OpenAI has adopted; it has resulted in some interesting behavior in which the LLM model begins responding to a prompt and then deletes the text when the classifier realizes that the generated output is potentially problematic.

Classifiers are a different type of machine learning model than LLMs. While they are also prediction based, classifiers should be much more reliable as they are effectively comparing the outputted text with text that was curated to include examples of different classes (i.e. problematic examples and nonproblematic examples) of output. Classifiers will have an important role in moderating LLMs but there are inherent limitations in relying on them. This includes the need for classifiers to be trained on specific examples, which first requires that all examples of problematic output are identified. Another important limitation is language; the training data for classifiers is more likely to come from the most widely spoken languages, including English, meaning that questions asked and answered in other languages may escape moderation. The bigger challenge with classifiers is that they are being trained on a proprietary basis. The producers of open-source LLMs are not developing or releasing classifiers to go with the LLMs to moderate their output. Work is needed to develop open-source classifiers that can be built into LLM pipelines to moderate potentially problematic output. The challenge, of course, is that absent a requirement to do so, a developer could simply decide to use the LLM without also using the classifier.

A final consideration around mitigation is that there may be cases in which non-moderated output should be provided. Moderation by its nature will result in some benign or even positive use cases being prevented. In the CBRN domain, this can mean that LLMs often will not help nonproliferation researchers analyze text about a specific proliferation issue, for example. One solution to this is to allow certain groups of users to bypass the classifier and interact directly with the unmoderated LLM. Careful examination of when and under what circumstances such access should be granted is needed. Any such access would likely have to be carefully controlled based on organizational and functional needs, while also accounting for other issues such as export controls which might create a need to restrict output based on nationality.

Conclusions

This paper is a first attempt to set out how LLMs could assist in the production of CBRN and how such risks can be mitigated. It may not be immediately apparent how the text-based outputs of an LLM could help in producing weapons of mass destruction, but there are certain areas where LLMs may provide assistance to nefarious actors, going beyond the information available on the web and in books. Of particular relevance to CBRN production processes, given that LLMs have been trained on vast quantities of computer code and their ability to generate scripts and code, they could assist in engineering design and computer simulation relevant to specific CBRN production steps. This role may expand in the future as fields such as generative design and advanced manufacturing capabilities grow.

Given this, it is necessary to develop mitigation measures to prevent LLMs from producing problematic CBRN related output. This paper argues that such measures should be reasonable and proportionate and has identified several built-in and proactive mitigation measures. In addition to this, the paper has also raised the need to examine situations which should not require mitigation measures. In some cases, companies should avoid restricting the ability of LLMs from producing useful outputs for all users that could make significant positive contribution in fields that involve dual-use technologies.

www.nonproliferation.org/dc

nonproliferation.org

Middlebury Institute *of*
International Studies at Monterey
James Martin Center for Nonproliferation Studies