This risk assessment focuses on cybersecurity threats that could arise from the open source release of a large language model (LLM). We will follow the general quantitative risk assessment methodology outlined by SaferAI, with a focus on quantifying the risks associated with different cyber attack vectors enabled by the LLM.

We will make the simplifying assumption that the LLM is released open source without any restrictions. This removes the complicating factors of analyzing access control policies.

The key steps in the risk assessment process are:

1. Identify potential cybersecurity attack vectors arising from the LLM
2. Assess the likelihood of each attack vector being exploited
3. Estimate the potential impacts/harms from each attack vector
4. Quantitatively combine the likelihood and impact estimates into an overall risk score for each attack vector
5. Identify risk mitigation strategies to lower the risk scores

Attack Vectors
We first brainstorm the potential ways the open source LLM could be used to execute or enable cyber attacks:

- Spear Phishing content generation - Using the LLM to dynamically generate personalized/contextualized phishing emails or messages to improve social engineering attacks. This could enable more effective credential theft, installation of malware, etc.

- Automated vulnerability discovery - The LLM could be fine-tuned to analyze software/systems and identify potential vulnerabilities like buffer overflows, logic errors, etc. This would significantly lower the barrier for attackers looking to hack systems.

- Automated exploit generation - Building on vulnerability discovery, the LLM could potentially generate functional exploits like injectable shellcode tailored to discovered flaws. Further lowering barriers to system compromise.

- Misinformation content generation - Generating fake news, propaganda, etc. could be used to manipulate perceptions and opinions for political ends. This has a cyber element in disseminating the misinfo via social media, fake accounts, etc.

There are likely more potential attack vectors, but these cover a broad enough scope for analysis.

Likelihood Assessment
For each attack vector, we need to estimate the likelihood of it being successfully exploited in practice. This requires making judgments about the current capabilities of LLMs, the skills/resources of potential attackers, etc.

For example, spear phishing content generation seems moderately likely - LLMs can already generate passable human language, but may still need aid in crafting truly convincing context-specific content. Automated exploit generation seems less likely currently, as code generation abilities are not as advanced.

After deliberation, example likelihood estimates on a 5 point scale could be:

- Spear Phishing content generation: Likely (4/5)
- Automated vulnerability discovery: Possible (3/5)
- Automated exploit generation: Unlikely (2/5)
- Misinformation content generation: Likely (4/5)

Impact Assessment
Next we estimate the potential impacts if each attack vector is successfully exploited. For cyberattacks, relevant impact factors include:

- Financial/data loss
- System downtime
- Reputational damage
- Physical harms (if cyber-physical systems compromised)
- Loss of confidence in digital systems

The impact will depend on the scale and criticality of systems compromised. Qualitative ratings could be:

- Spear Phishing: Major (compromise of sensitive systems/data)
- Vulnerability discovery: Moderate (enables future attacks)
- Exploit generation: Major (easy compromise of systems)
- Misinformation: Moderate (erodes trust in institutions)

We would want to try quantifying impact in terms of dollars, repair time, etc. But qualitative ratings may be the best we can do initially.

Risk Scores
We can now combine the likelihood and impact estimates into overall risk scores for each attack vector:

- Spear Phishing content generation: Likely x Major = High Risk
- Automated vulnerability discovery: Possible x Moderate = Moderate Risk
- Automated exploit generation: Unlikely x Major = Moderate Risk
- Misinformation content generation: Likely x Moderate = Moderate Risk

This provides an initial risk profiling and prioritization for mitigations.

Risk Mitigation Strategies

Some potential mitigations to lower risks:

- Release only a scaled-down version of the LLM (limits capabilities)
- Obfuscate parts of the model (increases rediscovery difficulty)
- Watermark training data (enables tracing generated content)
- Limit sharing on hacker forums (reduces access to bad actors)
- Multi-factor authentication (protects against credential theft)
- Anomaly detection systems (identify unusual activity)

We would analyze each mitigation option using the risk assessment framework to estimate their risk reduction potential. This iterative analysis allows focusing resources on the most impactful options.

Conclusion

In this hypothetical example, we walked through a quantitative cyber risk assessment enabled by the open source release of an LLM. The methodology provides a systematic way to analyze and prioritize cyber threats arising from AI systems. There are still challenges in credibly estimating likelihoods and impacts that require expertise and data. But even initial qualitative applications of the framework can highlight risks and mitigations for further study. With iteration and scrutiny, we believe this methodology can become a valuable tool for AI safety.