The short answer is: it's probably not possible right now, it might be in the future.

A novice in a field who has access to LLMs can sometimes accomplish things that would normally require expert knowledge. LLMs can provide the user with information they wouldn't even know how to find and, when they're putting this info into practice, help troubleshoot their actions. This has led some people to worry that current or future LLMs might be instrumental in enabling a terrorist group or other small bad actor to cause large-scale harm using methods that would otherwise require deep expertise — in particular, biological weapons.[1]

Most frontier LLMs released by major labs have been trained to not answer questions that could help users cause harm. However, proprietary LLMs can be jailbroken, and open-weights LLMs can be fine-tuned to remove their restrictions.[2] With that in mind, the key question becomes: how likely is it that a non-specialized actor with access to an unrestricted LLM would be able to build a biological weapon with a reasonable budget? Furthermore, how much does it help to have access to the LLM instead of just the internet?

The argument for LLMs being more useful than a search engine here is not that an LLM has access to more information, but that it can synthesize it more helpfully. For instance, a search engine might return information on how to procure sensitive materials, but cannot explain how to combine them in a specific setting, or design a plan to avoid raising suspicion. An LLM can critique plans and let the user iterate on them, as well as help overcome some types of roadblocks where a search engine would not be useful.[3]

As of 2025, LLMs seem unlikely to be helpful enough to enable such a plot, because they lack the skills needed to be an effective mentor. Such a mentor would have to be able to determine what you know and what you don't, find the biggest roadblocks, and redirect your focus when needed, and current LLMs arguably cannot do these things effectively. But people with relevant expertise disagree about whether LLMs might soon be sufficiently capable for this threat to become realistic. Here are a few and their thoughts:[4]

- Jonas Sandbrink (2023) is concerned.
- Kevin Esvelt (2023) is also concerned.
- A RAND report (2024) says it's unlikely with current models.
- An OpenAI report (2024) also says it's unlikely, but Gary Marcus is unconvinced.
- Luca Righetti thinks current models (as of 2024) are OK but we need better tests in the future, and is not convinced by OpenAI's o1 report.

---

[1] Other possible vectors for harm include hacking or large scale disinformation or extortion.
[2] This allows skilled users to bypass the tendency for LLMs to refuse potentially harmful requests which have been trained into them. This can also be done to a certain extent with closed-weights models.
[3] For instance, a question like the following is context-specific and might not have been answered on the web before: "I want to send my DNA samples to a lab that will combine them. How do I choose a lab that won't ask too many questions?"
[4] There was one attack on US soil where the perpetrator is known to have used an LLM to plan his actions, although all of the information he got from it could easily have been obtained by using conventional search.

- Anthropic's Claude 4 system card says an uncensored version substantially helps in planning such a plot, but not enough to reliably succeed.

Alternative phrasings

-

Related

- Might someone use AI to destroy human civilization?
- What about people misusing AI?

Scratchpad

Mikhail says we should emphasize that there is a barrier to entry to bioterrorism