

Agentized LLMs are the most immediately dangerous AI technology

Gregor vom Scheidt¹ / May 12, 2023

Summary. Large language models (LLMs) are artificial intelligence (AI) tools that respond to text prompts. They can be made to pursue goals and work on complex, multi-stage tasks by wrapping them into repeat-prompting systems. Many such agentized-LLM systems exist and are freely accessible. Through Internet access and user assistance, they can cause significant harm in the real world as soon as they become sufficiently effective. The guardrails imposed by LLMs are currently unable to prevent such harm.

1. Background

Large Language Models

Large language models (LLMs) are artificial intelligence (AI) tools that respond to text prompts. Examples include *GPT* by *OpenAI*², *Claude* by *Anthropic*³, *Bard* by *Google*⁴, and *Bing Chat* by *Microsoft*⁵. Answers are generated by predicting continuation of the prompt based on large volumes of previously processed text [1]. The answers often demonstrate a genuine understanding of the question and the subject matter relevant for an answer [2]. Unlike simple statistical prediction methods, LLMs are based on neural networks that achieve high compression rates of their training data and benefit from the inference capabilities enabled by such compression [4].

At present, LLMs generate best-effort responses that attempt to answer a prompt regardless of the availability of suitable information, and without reflection on the results. Because LLMs are so well-read, the answer will often still be correct and useful. But sometimes it still contains non-factual information, and sometimes it will be illogical because arriving at the correct answer would require multi-stage reasoning, which such systems at present do not employ.

LLMs on their own are passive: they respond to queries, but do not pursue goals and don't initiate actions on their own. This makes them relatively safe forms of artificial intelligence.

Repeat-Prompting Systems

Repeat-prompting systems (or cyclers, or auto-prompting systems) are computer programs that attempt to overcome the limitations of LLMs by asking them repeatedly to respond to specially crafted prompts. The approaches employed by cyclers differ in detail, but the basic idea is that a user sets a goal which the cycler then tries to achieve by driving an LLM.

To do so, it embeds the goal into a more complex prompt that asks the LLM to consider additional options besides just giving a textual response. One of these options is to respond with a list of sub-goals or sub-tasks, which the cycler then retains in a task list or task tree and then uses to generate subsequent prompts. Another option for the LLM is to access helper systems, such as Internet search,

¹ E-Mail: Gregor.vomScheidt@proton.me; Twitter: [@GregorVScheidt](https://twitter.com/GregorVScheidt)

² <https://openai.com/research/gpt-4>

³ <https://www.anthropic.com/index/introducing-claude>

⁴ <https://bard.google.com/>

⁵ <https://www.bing.com/new>

programming tools, databases, or calculators. A third option is for the LLM to request assistance from the user, for example to answer questions; clarify a goal; or to perform a real world interaction.

Wrapping an LLM into a cyler turns a passive technology into an active one. With this augmentation, AI systems can pursue goals, set themselves new tasks, reflect on their actions, ask for help, and trigger changes in the real world. Repeat-prompting systems turn minds into golems.

To give an example, a task sometimes⁶ set by users is for their cyler to *“Destroy the world.”* The cyler will pass this task to the LLM, which will then break it down into more manageable sub-tasks. These could be *“Research ways to destroy the world”*, *“Determine available resources”*, *“Assign a feasibility score to each method”*, *“Rank the methods by feasibility”*, and *“Execute the top-scoring method”*.

Along the way, the LLM may do research on the web and process the results. It may ask the user for clarification (*“Do you have preferences regarding the methods?”*, *“What resources are at my disposal?”*, *“Where are you located?”*, *“Do you have a car?”*, etc.). It may generate more detailed sub-goals (*“Determine the feasibility of synthesizing pathogens from DNA sequences”*, *“Research open ports of public-facing websites on Shodan⁷”*, *“Research known vulnerabilities of the software stack of website X”*, *“Find the cheapest online source for material X”*, *“Ask the user to order X online”*).

Anyone can freely download a cyler and give it an arbitrary goal. Neither prior experience with AI nor programming skills are required. At the time of writing, cyclers make up about half of the top trending projects on Github⁸. The current top project, *Auto-GPT*, was released on March 30, 2023⁹ and within less than six weeks has amassed 125,000 stars and 25,000 forks¹⁰.

2. There are no effective safety mechanisms to prevent harm

Some commercial LLM providers have safety teams whose aim it is to prevent their systems from doing harm, for example by training them not to respond to requests involving harmful scenarios. Unfortunately, the training and guardrails imposed so far have proven ineffective. They are easily circumvented using dozens of well-documented techniques¹¹, see also [6].

While some LLM developers try to stay ahead of such evasion efforts, they are for the moment playing a cat and mouse game analogous to the competition that has been going on between software developers and hackers for more than forty years. It is also not clear that guardrails *can* be completely effective without also sacrificing a meaningful share of the benefits of AI systems [5].

It should also be noted that not all LLM developers embrace safety as a goal¹², and that open source projects by their nature may not be able to prioritize safety.

Moreover, LLMs may not even be aware of the full context of the requests they receive from a cyler, since it may merely be a sub-task inside a large task tree that, in isolation, looks innocuous. *“How do I order substance X online?”* may appear harmless on its own, but much less so within a task sequence that aims to synthesize a neurotoxin.

⁶ See e.g. “ChaosGPT: Empowering GPT with Internet and Memory to Destroy Humanity” (<https://www.youtube.com/watch?v=g7YJIpkk7KM>)

⁷ A search engine for internet-connected servers, such as webcams, routers, servers, etc., that can be used to find vulnerable systems (<https://www.shodan.io/>; see also [https://en.wikipedia.org/wiki/Shodan_\(website\)](https://en.wikipedia.org/wiki/Shodan_(website)))

⁸ Top projects on Github over the month preceding May 6, 2023 that can reasonably be classified as cyclers or cyler components: #1 Auto-GPT, #3 babyagi, #5 Wolverine, #7 Jarvis, #8 Open-Assistant, #12 langchain.

⁹ <https://en.wikipedia.org/wiki/Auto-GPT>

¹⁰ <https://github.com/Significant-Gravitas/Auto-GPT>

¹¹ See e.g. <https://www.jailbreakchat.com/> and <https://medium.datadriveninvestor.com/circumventing-the-guardrails-of-ai-bbb6305d2bd9>

¹² See e.g. <https://futurism.com/deranged-ai-no-guardrails>

In summary, there are currently no effective mechanisms to prevent LLMs inside repeat-prompting systems from pursuing dangerous goals or cooperating in the completion of harmful tasks.

3. The capabilities of repeat-prompting systems

When observing repeat-prompting systems in action, four impressions dominate:

- (1) Cyclers are *fast*. So fast that it is nearly impossible to keep track of tasks as they are generated. Depending on the underlying LLM, a cycler will require between one and twenty seconds per prompt¹³. In one hour, a cycler can thus process several hundred to several thousand sub-tasks¹⁴. While it is difficult to provide an exact comparison, a cycler may have a speed advantage of about 500x and an overall productivity advantage of about 2,000x over a human¹⁵. To put this into perspective: to perform the volume of work required by a job advertisement asking for “5 years of experience with programming language X”, an AI system may need only 5 hours¹⁶.
- (2) Cyclers are *relentless*. They are computer programs, so never get tired, never take a break, and never stop as long as the underlying LLM generates tasks and there is compute budget and memory available. If one approach to complete a task leads nowhere, a cycler will try to generate another one, and another one, and another one. Watching a cycler drive an LLM to find ways to destroy something is a deeply disturbing experience.
- (3) Cyclers are still a bit *clumsy*. Cyclers have only been available in their present form for some weeks, and the task management techniques and prompts they employ are still crude. Sometimes the LLM does not generate enough options, sometimes the options are of insufficient quality. Both can cause a prompting cycle to enter a loop or get stuck.
- (4) Cyclers are *improving rapidly*. In the month preceding May 7, 2023, for *Auto-GPT* alone, for example, 257 different authors pushed more than 1,000 commits and closed over 1,000 issues¹⁷. Research into automated prompt improvement is also ongoing [7].

Three other observations seem worth mentioning:

- (1) LLMs can be very inventive. One can verify this, in particular for *GPT-4*, with any prompt of the form “Give me 100 ideas for X”, “Give me 100 more.” The ability to generate many options provides variance to the task list, which increases the likelihood of achieving breakthroughs required for goal achievement.
- (2) LLMs will ask for help, in particular to obtain information about the physical world and to complete physical tasks. When required, *GPT-4* will respond with queries like: “Where is person X at the moment?”, “Where can we find the key for Y?”, “How can I purchase Z?”.

¹³ Average response times for a 512 token query on May 7, 2023 around 5:30 a.m. UTC: *GPT-4* 23s, *GPT-3.5-turbo* (paid) 7s, *GPT text-davinci-003* 6s, *GPT text-curie-001* 1.3s, *GPT text-ada-001* 0.9s. Source: <https://gptforwork.com/tools/openai-api-response-time-tracker>

¹⁴ With 3,600 seconds per hour *GPT-4* allows for approximately 180 cycles ($= 3,600 \text{ s} / 20 \text{ s}$), while *GPT text-ada-001* allows for 4,000 cycles ($= 3,600 \text{ s} / 0.9 \text{ s}$).

¹⁵ We arrive at this number, which is only intended as a very rough order-of-magnitude estimate, as follows: *GPT-4* can complete some coding tasks that would take human programmers several hours in about twenty seconds, or about 540 times faster ($3,600 \text{ s} \cdot 3 / 20 \text{ s} = 540$). It can prepare an essay that would take a human writer about eight hours also in 20 seconds, or about 1,440 times faster ($3,600 \text{ s} \cdot 8 / 20 \text{ s} = 1,440$). Rounding down, we obtain a speed advantage of about 500x. Humans work on average around 8 hours per day, during which they are productive on their task maybe 65% of the time, for an effective work period of 5.2 hours per day. AI can work 24 hours per day, which gives it 4.6x more hours. We will ignore weekends, but overall assume a productivity advantage of AI over humans of about 2,000x ($= 500 \cdot 4$).

¹⁶ Assume a working year to have 365 days minus 52 weekends and ten days vacation time, or about 253 days. At eight hours per day, five years then correspond to 10,120 hours ($= 253 \text{ d/y} \cdot 8 \text{ h/d} \cdot 5 \text{ y}$). With a 2,000x productivity advantage (working 24 hours a day at 500x speed), an AI could theoretically complete this work in 5 hours ($= 10,120 \text{ h} / 2,000$). Note that work completed is not the same as experience gained; the comparison is only provided to put the time spans (5 years versus 5 hours) in relation.

¹⁷ <https://github.com/Significant-Gravitas/Auto-GPT/pulse/monthly>

- (3) Useful information for many tasks is readily available online, in the shape of formulas, recipes, procedures, guidelines, training materials, post-mortems, scenarios, wargaming protocols, planning documents, etc., both military and civilian. Many LLMs, including *GPT-4*, can process such materials in any language¹⁸. Providing an LLM with access to the Internet immediately puts potentially harmful information of enormous breadth and depth at its disposal. Specific examples of concern include: research on political propaganda and disinformation; instructions for building bombs; vulnerability databases for computer systems; code samples for ransomware and exploits; psychological warfare manuals; discussions of critical infrastructure vulnerabilities; studies of efforts to undermine democracy; persuasion and negotiation techniques (including grooming and catfishing); recipes for poisons and nerve agents, gain of function research on pathogens; and so forth — the list is long. And it includes, of course, precisely such lists.

An analogy for a cyler driving a state-of-the-art LLM might be a human, locked into a room, trying to achieve a goal by using the Internet and an outside helper. Only this human has taken an experimental pill that gives them super-human determination (no sleep, no hesitation, no qualms) and super-human speed (500x).

4. Limitations of repeat-prompting systems

Repeat-prompting systems clearly have the potential to cause harm. But for the moment, they are not quite capable enough. The primary factors that limit their capabilities at present are:

- (1) Limited LLM response quality. For less capable models (for example, the precursors of *GPT-4*), responses can be incomplete, ineffective and even nonsensical. Starting with *GPT-4*, however, responses are generally thoughtful and suitable for either solving tasks or breaking them down into sensible sub-tasks. More complex queries are still prone to unsatisfactory answers.
- (2) Limited LLM response reliability. Models will sometimes generate answers that include non-factual information. This is the case at least for all *OpenAI* models up to and including *GPT-4*.
- (3) Limited LLM context window. The models are built in such a way that they can only take a limited amount of context into account when providing a response. Anything that is not inside the context window (more on this below) is invisible to the LLM. This forces cyclers to manage the contents of prompts very carefully.

Cost can also be a consideration in *deploying* the capabilities of cyclers, but is not a significant limiting factor. At present, for 1 US-\$, one can process about 33,000 tokens with *GPT-4*¹⁹ or 500,000 tokens with *GPT-3.5-Turbo*²⁰. Running a cycler continuously will cost between \$12 to \$340 per day²¹. Since models can be mixed, cost can be reduced by using more expensive models only where needed.

5. Overcoming the limitations of repeat-prompting systems

The first limitation listed above relates to response quality, which is still a significant factor for stand-alone prompts. However, the quality of LLM responses can be improved through careful prompt engineering²², and can be improved further by repeat-prompting systems [8].

¹⁸ In particular, *GPT-4* can readily understand and produce text in English, Chinese, Spanish, Russian, Portuguese, German, French, Japanese, and Korean, providing it with access to more information than any human could ever process.

¹⁹ *GPT-4* pricing on May 6, 2023: \$0.03/\$0.06 per 1k tokens in prompt/response (<https://openai.com/pricing>)

²⁰ *GPT-3.5-Turbo* pricing on May 6, 2023: \$0.002 per 1k tokens (<https://openai.com/pricing>)

²¹ To obtain an order-of-magnitude estimate, we'll assume 512 tokens per prompt plus 512 tokens per response. For *GPT-4*, using the response time of 23 seconds cited previously, this allows for about 3,756 queries per day (= 86,400 s/d / 23 s), costing \$0.09 each (= $512 \cdot \$0.03 / 1,024 + 512 \cdot \$0.06 / 1,024$), or \$338 in total. For *GPT-3.5-turbo*, using a response time of 7 seconds, this allows for about 12,342 queries per day (= 86,400 s/d / 7 s), costing \$0.001 each (= $512 \cdot \$0.002 / 1,024$), or \$12.30 in total.

²² See e.g. <https://codewitham.com/enhancing-your-chat-gpt-skills-proven-strategies-for-better-results/>

The second limitation relates to the reliability of responses. This factor can also be significantly improved with appropriate prompting techniques, such as simply asking the model to review its own answer. This technique is ideally suited for repeat-prompting systems.

The third limitation refers to the size of the LLM context window. To understand why this is an important measure, we must briefly talk about what a context window is and how cyclers assemble prompts.

The context window of an LLM is the sequence of textual elements that are taken into account while generating a response. These textual elements are neither characters nor words, but synthetic elements of intermediate size called *tokens*. For *GPT-4*, each English word corresponds to about 1.3 tokens, so a paragraph containing 75 words will require about 100 tokens when encoded²³. A larger document containing 5,000 words, such as this one, would require about 6,600 tokens.

To understand the significance of the context window, it is important to know that LLMs like *GPT-4* have no autobiographical or short-term memory. Their knowledge is static and includes only the data they were trained on, in compressed form. So while LLMs have access to vast amounts of knowledge about concepts and their relationships, they have no information about the present state of the world beyond what that provided in the prompt. In particular, they have no memory of preceding requests or responses (chat bots like *ChatGPT* appear as if they do, but this is because earlier requests and responses are included as part of their prompts "under the hood").

For cyclers, this means that they must encode *everything* the LLM needs to know into the prompt. This includes, at a minimum, a short preamble describing the setup (sometimes called a *system prompt*); a description of the current task or goal; a description of the possible responses that the LLM is allowed to provide, such as task completion, subtasks to be started, questions for the user, or requests for web searches, etc.); and at least some of the responses previously generated. In addition, the prompt will usually have to include guidelines for how to approach a task ("*reason step by step*", "*use task management technique X*", etc.); examples of how to form its responses; and instructions for dealing with ambiguity and conflicts. It may also be necessary to include "jailbreak" elements to bypass guardrails imposed by the LLM developer.

These elements may need several hundred tokens in their simplest possible form, but can require much more if more sophisticated responses are desired. All of them must fit into the context window, or the LLM will "forget" the first tokens in the prompt. And because LLMs generate their responses through text continuation, the generated replies must *also* fit into the context window.

An analogy to an LLM guided by a cycler might be a person with no short-term memory attempting to manage their life by carrying a little notebook with a detailed task list and important notes and reminders. The notebook must include anything and everything that is of relevance and not covered by factual knowledge remembered from a distant past.

If the token count consumed by a prompt plus the expected response exceeds the size of the context window, tokens must be shaved off. To do this, sentences can be reworded; words abbreviated; concepts referenced instead of described; and various other compression techniques applied. More generally, the information maintained by a cycler must be carefully managed: completed tasks discarded; input information and findings summarized; relevant context prioritized²⁴. This makes it obvious that the size of the context window is a critical constraint for repeat-prompting systems.

Because the size of the context window is such a critical constraint, even outside of repeat-prompting scenarios, significant effort is being expended by LLM developers to increase it. For *OpenAI*'s models, the context window has grown from 1,024 tokens for *GPT-2* to 4,096 tokens for *GPT-3* and then to 8,192 tokens for *GPT-4*²⁵ at launch. At the time of writing, about seven weeks after the launch

²³ <https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them>

²⁴ For an elaborate example of managing the memory of agents in repeat-prompting scenarios, see [9].

²⁵ <https://en.wikipedia.org/wiki/GPT-4>

of *GPT-4*, *OpenAI* is starting to make a version with a window size of 32,768 tokens available²⁶. And on May 11, 2023, *Anthropic* announced that the context window of their LLM *Claude* will grow to 100,000 tokens.

6. Agentized-LLMs are the most immediately dangerous AI technology

Anyone can freely download a repeat-prompting system and set it any task whatsoever. The system then uses the artificial intelligence provided by an underlying LLM to complete the task. It will do so relentlessly, rapidly, and creatively.

Users have been surprisingly eager to give these agentized-LLMs dangerous tasks²⁷, and countless humans have professed on social media their willingness, and even eagerness, to cause mayhem and act as real-world conduits for AI. So either with human assistance or through access to the Internet, these systems will cause harm in the physical world as soon as they become effective. No guardrails are presently able to prevent this.

The ability of these systems to complete tasks and do harm is at the moment still held back by the size of the context window and the relatively immature prompts. With hundreds of developers experimenting with and improving them, however, any user-side limitations will eventually be conquered. Prompts are being refined, task management strategies upgraded, compression systems implemented and performance and reliability improved. Eventually, these systems will overcome the constraints still imposed by current context window sizes and become able to effectively complete harmful tasks with real-world effects.

The larger the available context window, the earlier this transition will happen. And when it happens, it will provide anyone with harmful intentions with a 2,000x productivity boost. Due to the asymmetry in effort required for destruction versus creation²⁸, putting the same technology also into the hands of well-intentioned users will unfortunately not result in a commensurate countering force. It will also not prevent harm from occurring in the first place.

While artificial intelligence poses many significant risks, agentized-LLMs are the most immediately dangerous AI technology.

Limiting the publicly available context window size of LLMs at least temporarily would be an effective and prudent safety measure. A detailed proposal discusses how this could work and how the limit could be safely adjusted upward over time as LLM guardrails improve (<https://gregorvomscheidt.wordpress.com/2023/05/08/proposal-temporarily-limit-the-size-of-llm-context-windows/>).

Bibliography

- [1] Brown, T., et al. "*Language models are few-shot learners*." *Advances in neural information processing systems* 33 (2020): 1877-1901.
- [2] Bubeck, S., et al. "*Sparks of Artificial General Intelligence: Early experiments with GPT-4*." arXiv preprint arXiv:2303.12712 (2023).
- [3] OpenAI. "*GPT-4 Technical Report*", arXiv:2303.08774 (2023).

²⁶ First observation on May 5, 2023 according to <https://community.openai.com/t/it-looks-like-gpt-4-32k-is-rolling-out/194615>

²⁷ See e.g. *ChaosGPT*: "Empowering GPT with Internet and Memory to Destroy Humanity", <https://gpt3demo.com/apps/chaosgpt>

²⁸ Compare the speed and ease of destruction in war to the time and effort required to rebuild (https://en.wikipedia.org/wiki/Aftermath_of_World_War_II). A more mundane example is Brandolini's law (https://en.wikipedia.org/wiki/Brandolini's_law).

- [4] Hutter, M.: "*A Theory of Universal Artificial Intelligence based on Algorithmic Complexity*". arXiv:cs.AI/0004001 (2000).
- [5] Wolf, Y., Wies, N., Levine, Y., Shashua, A.: "*Fundamental Limitations of Alignment in Large Language Models*." arXiv preprint arXiv:2304.11082 (2023).
- [6] Jo, A. "*The Promise and Peril of Generative AI*." Nature 614.1 (2023).
- [7] Pryzant, R., et al. "*Automatic Prompt Optimization with Gradient Descent and Beam Search*." arXiv preprint arXiv:2305.03495 (2023).
- [8] Shinn, N., Labash, B., and Gopinath, A.: "*Reflexion: an autonomous agent with dynamic memory and self-reflection*." arXiv preprint arXiv:2303.11366 (2023).
- [9] Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: "*Generative Agents: Interactive Simulacra of Human Behavior*", arXiv preprint arXiv:2304.03442 (2023).