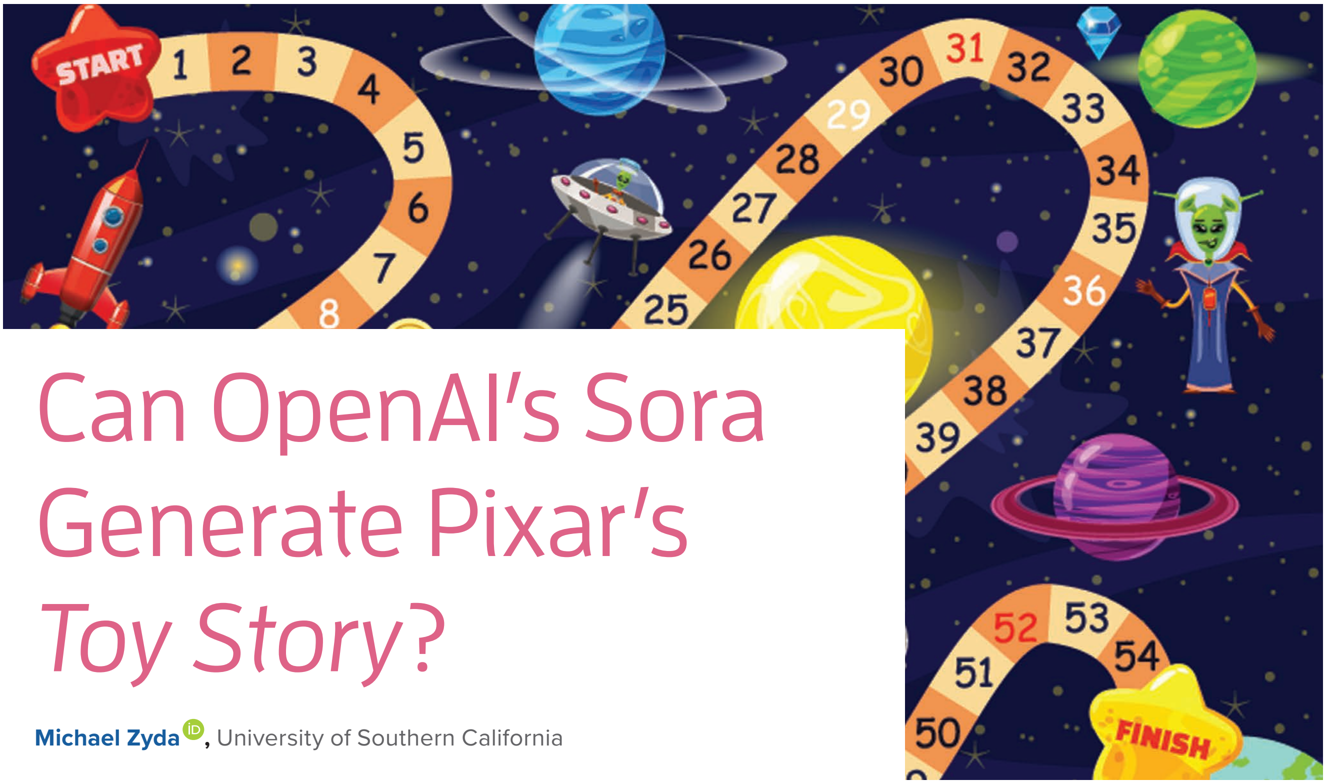
GAMES



Can OpenAI's Sora Generate Pixar's *Toy Story*?

Michael Zyda, University of Southern California

In a previous "Games" column, I raised the question of whether OpenAI's Sora could generate the complete Pixar animated film Toy Story. *We discuss how to think about the computational requirements of such an endeavor.*

In a previous "Games" column,[1] I spoke about Sora's ability to generate short videos of 30 s or less in length, videos of detailed "photorealistic scenes." The longest video released at that time was 17 s and was generated using a 38-word text prompt input to Sora, built on top of OpenAI's foundational model GPT.[4] The text prompt used for the 17-s video is the following:

> "Beautiful, snowy Tokyo city is bustling. The camera moves through the bustling city street, following several people enjoying the beautiful snowy weather and shopping at nearby stalls. Gorgeous sakura petals are flying through the wind along with snowflakes."[2]

At the time of writing, I suggested that I would be more impressed if we could hand Sora the text prompt that would generate the complete computer-generated Pixar film *Toy Story*, an 81-min-long film (Figure 1). Before we go into that, we need to analyze what we can get out of the Tokyo scene video.

How much time did Sora require to generate the 17-s Tokyo scene video?

Levy[2] said the following:

> "The researchers I spoke to won't say how long it takes to render all that video, but when pressed, they described it as more in the "going out for a burrito" ballpark than "taking a few days off." If the hand-picked examples I saw are to be believed, the effort is worth it."[2]

So there is maybe 30 to 45 min of computation for the 17-s video of the Tokyo scene, depending on how far away

Digital Object Identifier 10.1109/MC.2024.3393249
Date of current version: 26 June 2024

EDITOR MICHAEL ZYDA
University of Southern California;
zyda@mikezyda.com

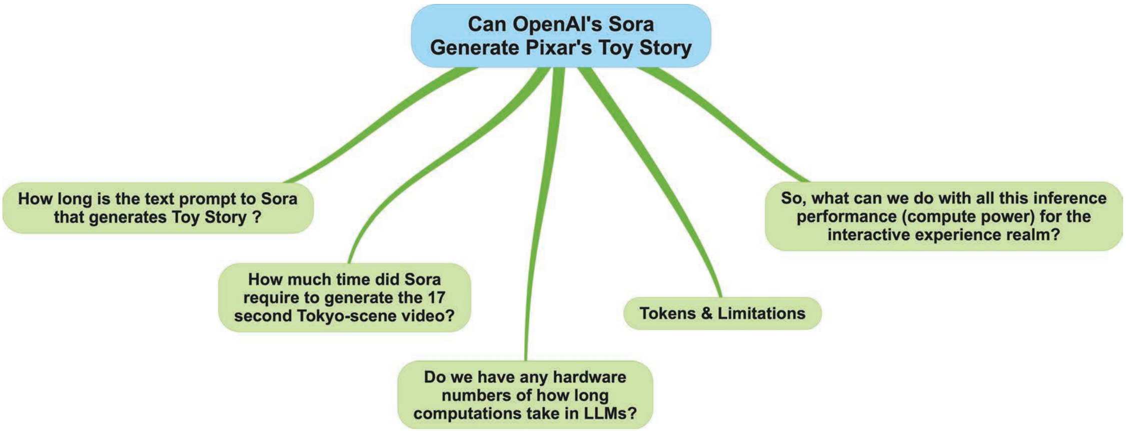

Can OpenAI's Sora Generate Pixar's Toy Story
How long is the text prompt to Sora that generates Toy Story ?
How much time did Sora require to generate the 17 second Tokyo-scene video?
Do we have any hardware numbers of how long computations take in LLMs?
Tokens & Limitations
So, what can we do with all this inference performance (compute power) for the interactive experience realm?

FIGURE 1. The organization of this column.

the burrito restaurant is and the length of time required for consumption. I am going to use 45 min for the creation of the 17-s Tokyo scene video.

Another thing that is important to note is that there is no real story in the Tokyo scene video. It is just a brief walk down a block of virtual terrain. This means that when we create the input prompt for *Toy Story*, we need to include prompt information for the 3D world of *Toy Story* plus the actual script for the story.

Do we have any hardware numbers of how long computations take in large language models (LLMs)?

There is a very nice discussion on computational performance of NVIDIA's A100 GPU:

> "Another source claims that a single NVIDIA A100 GPU can run a 3-billion parameter model in about 6ms. Considering this speed as the reference point, a single NVIDIA A100 GPU could take 350ms seconds to print out just a single word on ChatGPT. Since the latest version, ChatGPT-3.5, has over 175 billion parameters, it will need around five A100 GPUs to perform the necessary action to get an output for a single query. The average number of at least 8 A100 GPUs is derived after taking ChatGPT's capability of outputting around 15-20 words per second."[5]

Now, the A100 GPU (1×) came out in 2021 and has been replaced by the H100 AI GPU (11×), the H200 AI GPU (18×), and now the Blackwell (44× and US$30,000 to US$40,000), but the A100 is still widely used for LLMs, primarily because the A100 was very expensive, about US$10,000 for one, and the serious users of these have purchased tens of thousands, some 20,000 or more A100s (Figure 2).[8]

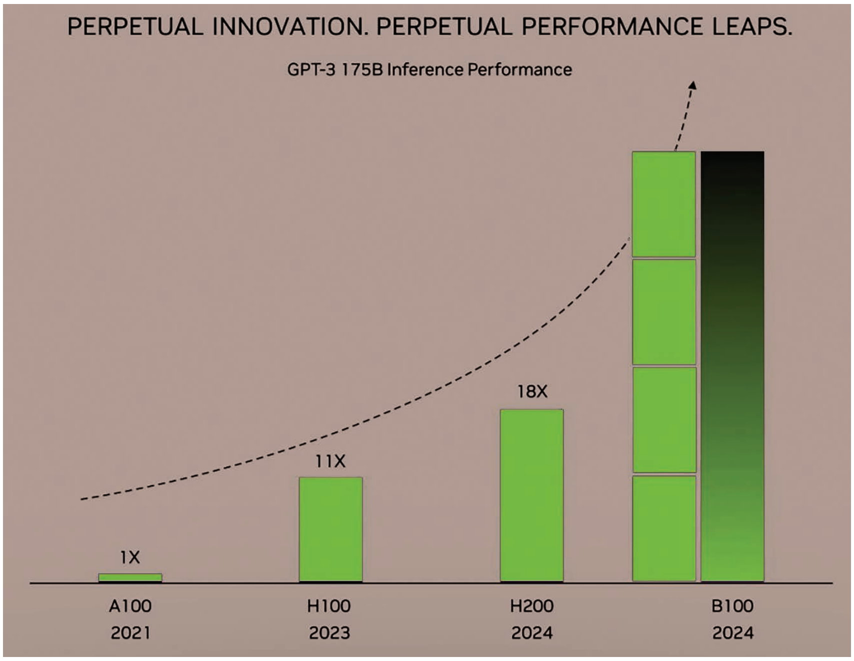


FIGURE 2. The NVIDIA B100 AI GPU against A100, H100, and H200 in AI inferencing. (Source: NVIDIA and also Garreffa.[8])

We are going to assume that the Tokyo scene video was computed on A100 NVIDIA hardware. If we had the ability to recompute it using B100 at 44× the speed, then we would see the Tokyo scene video come out in just over one minute of time. Not enough time to go out for a burrito.

TOY STORY

Now, *Toy Story* is an 81-min-long film (4,860 s). We can do a rough computation on how many words minimum we need for the *Toy Story* input prompt to the GPT model (Figure 3). If we take 4,860 s and divide it by the 17 s (286 times more words required) of the Tokyo scene video and then multiply that (286) by the 38 words of the Tokyo scene video, we then get 10,868 words minimum for the input prompt we need to give to the GPT model. If we run this on A100 hardware, this would be 286 times 45 min of computation required,12,870 min or 214.5 h for this minimum input prompt for *Toy Story*. If we run this on B100 hardware, 44× faster, then this is 292.5 min or 4.875 h, somewhat more tractable.

Now, the screenplay for *Toy Story* is actually online and can be downloaded.[11] That screenplay has 21,455 words in it according to Microsoft Word, so we are not too far off, but the screenplay just tells the story and probably not enough description to actually get the art style we all love in the actual film.

If we use that screenplay as the basis for similar computations to those performed above and again utilize A100 hardware, the computation time is 25,425 min or 423.75 h. When the computer graphics field of radiosity was first starting out, radiosity had similar computation time requirements: they called it *geological time* in honor of how slow things move in that sphere of endeavor. If we use B100 hardware, again 44× faster, then we need 578 min or 9.63 h for our computation. So not burrito time but "run this, go home and sleep and come back tomorrow" time. And that is just to get one run of the video, and I am assuming that every single tweak we want to make/test, that is what we would have to do. And remember I said something about the art style needing to be specified, and that is not in my computation.

TOKENS AND LIMITATIONS

One of the most trying things to understand is the input limitations of OpenAI's GPT models: Sora is built on top of GPT. This means there is a maximum size of text prompt we can provide to the GPT model. The limitation in size of input prompts is not specified by word or character count but rather by the number of tokens.[9,10] Tokens are defined as "groups of characters, which sometimes align with words, but not always. In particular, it depends on the number of characters and includes punctuation signs or emojis."[9]

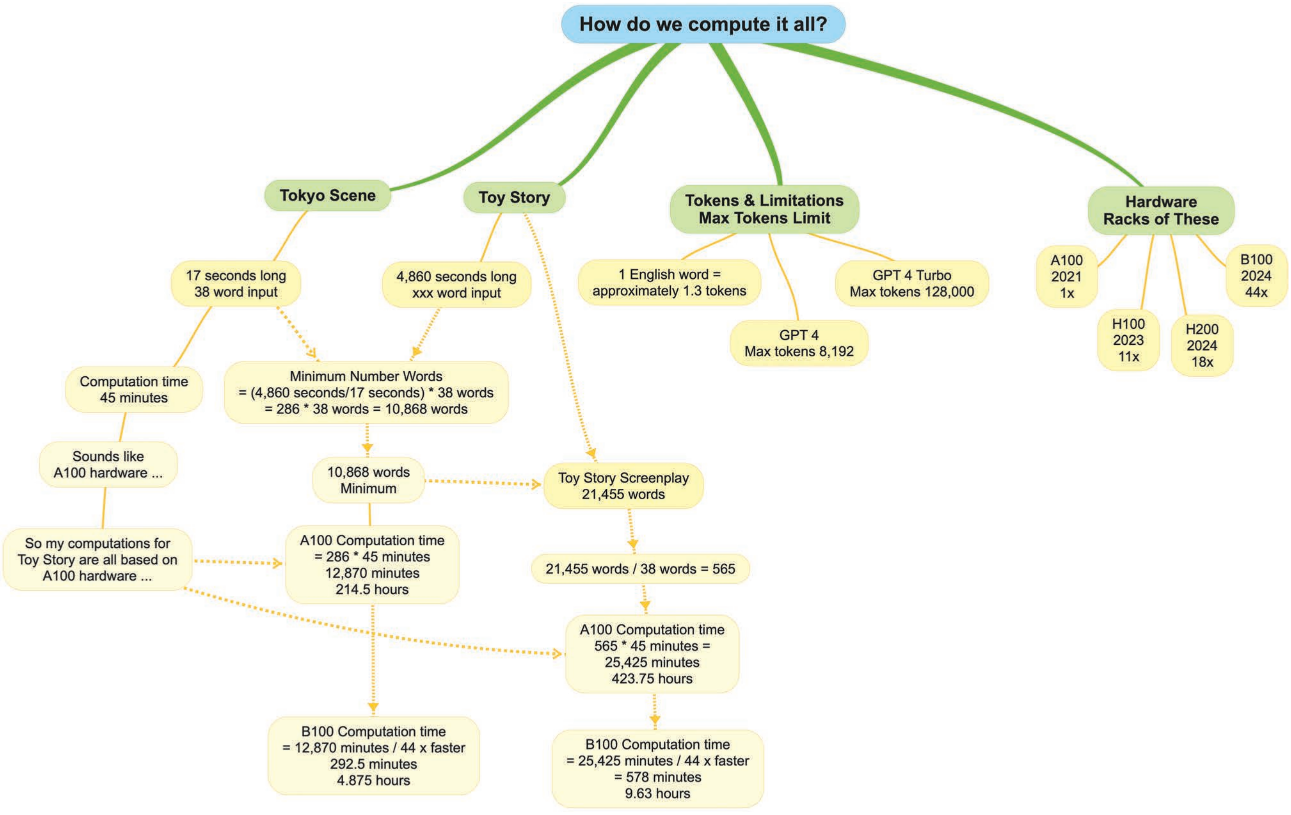


FIGURE 3. How do we compute it all?

This basically means you only find out by trying your prompt. My first prompt was to hand ChatGPT a PDF of a screenplay, and it just balked at it and did not provide any meaningful error message. Chat GPT-4 has 8,192 tokens as the maximum input, about 5,734 words using the rule of thumb of one English word equal to approximately being 1.3 tokens.[9] A complete list of the token max for each of OpenAI's GPT models is here.[10] GPT-4 has a maximum number of tokens of 8,192, and GPT-4 Turbo has a maximum number of tokens of 128,000. For us to compute the complete *Toy Story*, we will need to use GPT-4 Turbo or better.

So, what can we do with all this inference performance for the interactive experience realm?

Well, let's assume that NVIDIA Blackwell farms become prevalent and available everywhere (and we have enough money and power to utilize them). Let's assume that we can make the running of LLM models more parallel and that we have all the fast memory we require. Let's assume that we know how to author input prompts to our LLMs that can provide us the 81 min or more of content in either filmed entertainment form or interactive game form. Maybe we can create films or episodes of shows where we can name the characters and have them look like family members or other actors, virtual or animated, that are also beloved. I know my granddaughter would love to be able to put herself into a show like *Bluey* and would maybe even like to specify the storyline or part of the storyline. We are not far away from the ability to make this happen. It just might be expensive in terms of inference compute power. For commercial film production, this may not be a problem as long as the films being made are blockbuster audience sized and can be paid for by tickets sold.

For interactive use, we need to be producing 60 to 90 to 120 frames per second of animated content. We need those worlds produced to be interactive such that we can reach out and touch 3D objects. The inference computation required is much larger than that required for film and show production. The additional thing to think about is that the large computation infrastructure is not in someone's home but off on the Internet in a server facility. This means that the delivery of real-time interactive worlds will have networking requirements and networking problems similar to those of Google's Stadia. We are not close to utilizing generative AI models in interactive fashion unless those model computations are inside of the device we are playing on. Maybe if Apple ever figures out how to get into the AI business, we will see this inside of their iPhones and iPads as they have the right processor architecture for something like a small language model. Maybe one of my next columns will be about the importance of networking for the interactive game space and perhaps a column on how to draft meaningful input text prompts for processing by LLMs like Sora.

> Maybe we can create films or episodes of shows where we can name the characters and have them look like family members or other actors, virtual or animated, that are also beloved.

COMMENTS?

If you have comments about this article, or topics or references I should have cited or you want to rant back to me on why what I say is nonsense, I want to hear. Every time we finish one of these columns, and it goes to print, what I'm going to do is get it up online and maybe point to it at my Facebook (mikezyda) and my LinkedIn (mikezyda) pages so that I can receive comments from you. Maybe we'll react to some of those comments in future columns or online to enlighten you in real time! This is the "Games" column. You have a wonderful day.

ACKNOWLEDGMENT

I wish to thank all of the AI and generative AI researchers that are connected to me on LinkedIn and Facebook; their online presences, commentaries, and latest published papers are greatly appreciated!

REFERENCES

1. M. Zyda, "Today generative AI is just a parlor trick," *Computer*, vol. 57, no. 5, pp. 98–101, May 2024.
2. S. Levy, "OpenAI's Sora turns AI prompts into photorealistic videos," *Wired*, Feb. 15, 2024. [Online]. Available: https://apple.news/A4rkDNyHGQgmuYotzVIOazQ
3. M. Zyda, "Large language models & generative AI, Oh My!" *Computer*, vol. 57, no. 3, pp. 127–132, Mar. 2024.
4. Raj. "How to use OpenAI Sora?" Medium. Accessed: Apr. 24, 2024. [Online]. Available: https://medium.com/@ecommerce_plan/how-to-use-openai-sora-41218d9d6142#:~:text=Technology%20behind%20the%20scenes%3A,to%20match%20the%20text%20prompt
5. "OpenAI's ChatGPT reportedly costs $100,000 a day to run." CIO Coverage. Accessed: Apr. 24, 2024. [Online]. Available: https://www.ciocoverage.com/openais-chatgpt-reportedly-costs-100000-a-day-to-run/#:~:text=Considering%20this%20speed%20as%20the,output%20for%20a%20single%20query
6. K. Leswing. "Meet the $10,000 Nvidia chip powering the race for

A.I." CNBC. Accessed: Apr. 24, 2024. [Online]. Available: https://www.cnbc.com/2023/02/23/nvidias-a100-is-the-10000-chip-powering-the-race-for-ai-.html

7. By M. A. Cherney, S. Nellis, and M. Singh. "Nvidia stock climbs as CFO says new chip to ship in 2024." Reuters. Accessed: Apr. 24, 2024. [Online]. Available: https://www.reuters.com/technology/red-hot-nvidia-dips-after-it-unveils-new-ai-chip-2024-03-19/#:~:text=Called%20Blackwell%2C%20Nvidia’s%20new%20processor,firms%20that%20use%20its%20technology

8. A. Garreffa. "NVIDIA officially teases next-gen B100 Blackwell GPU: Over 4x as fast as H100 AI GPU." TweakTown. Accessed: Apr. 24, 2024. [Online]. Available: https://www.tweaktown.com/news/94338/nvidia-officially-teases-next-gen-b100-blackwell-gpu-over-4x-as-fast-h100-ai/index.html

9. How to use OpenAI GPT tokens? GPT for Work. [Online]. Available: https://gptforwork.com/guides/openai-gpt3-tokens#

10. 'Models – Discussion on number of tokens max in each model." Open AI. Accessed: Apr. 24, 2024. [Online]. Available: https://platform.openai.com/docs/models/overview

11. "Toy story script." Daily Script. Accessed: Apr. 24, 2024. [Online]. Available: https://www.dailyscript.com/scripts/toy_story.html

MICHAEL ZYDA is the founding director of the Computer Science Games Program and a professor emeritus of engineering practice in the Department of Computer Science, University of Southern California, Los Angeles, CA 90089 USA. Contact him at zyda@mikezyda.com.

Digital Object Identifier 10.1109/MC.2024.3408588