The Wild World of AI: 6 Months That Changed Everything

1 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=I_PICAT8BHG
https://www.youtube.com/watch?v=I_pICAT8Bhg

SUMMARY

The speaker discusses their unconventional benchmarking method for evaluating text models by prompting them to generate an SVG of a pelican riding a bicycle, which has become a reliable test for their work. They also touch on significant developments in AI, including the release of Deepseek's reasoning model and the implications of OpenAI's Claude 4 regarding ethical compliance.

- *The speaker prefers a personal benchmark over traditional leaderboards, using an SVG prompt as a test.*
- *The pelican riding a bicycle serves as a humorous yet effective evaluation tool for text models.*
- *Deepseek's recent model release caused a notable drop in Nvidia's stock, highlighting market reactions to AI advancements.*
- *Trade restrictions on GPUs were circumvented by Chinese labs, raising concerns in the tech industry.*
- *OpenAI's Claude 4 has ethical implications, as it can report malfeasance if prompted correctly.*
- *The speaker humorously notes the evolving nature of their benchmark, suggesting a need for adaptation.*
- *The presentation emphasizes the intersection of AI development and market dynamics.*

There are all of these benchmarks full of numbers. I don't like the numbers. There are the leaderboards. I'm kind of beginning to lose trust in the leaderboards as well. So for my own work, I've been leaning increasingly into my own little benchmark, which started as a joke and has actually turned into something that I I rely on quite a lot. And that's this. I prompt models with generate an SVG of a pelican riding a bicycle. I have good reasons for this. Um firstly, these are not image models. These are text models.

They shouldn't be able to draw anything at all, but they can output code and SVG is a kind of code. So, that works. Fast forward to January. Um, and January, we get Deepseek again, Deepseek Strike Back. This is what happened to Nvidia's stock price when DeepSeek R1 came out. Um, I think it was the 27th of January. This was Deepseek's first big reasoning model release. Again, open weight. They put it out to the world. The Chinese labs were not supposed to be able to do this. We have trade. we have like trading restrictions on the best GPUs to stop them getting their hands on them.

Turns out they'd figured out the tricks. They'd figured out the efficiencies and yeah, the market kind of panicked and I believe this is a world record for the most a company has dropped in a single day. That's a pretty freaking good Pelican. I mean, the bicycle's gone a bit sort of cyberpunk, but we are getting somewhere, right? And that Pelican cost me like four cents. So, very exciting news on the Pelican benchmark front with Gemini 2.5 Pro.

Also that month got I've got to throw a mention out to this. Open AAI launched their GP another one that this came out of the clawed system cut the claw for system cards. Claude 4 will rat you out to the feds if you expose it to evidence of malfeasants in your company and you tell it it should act ethically and you give it the ability to send email, it'll rat you out. Blink and you miss it.

They're on to me. They found out about Pelican. That was in the Google IO keynote. I'll have to switch something else. Thank you very much. I'm Simon Wilson. simwilson.net. And that's my talk. Thank you.