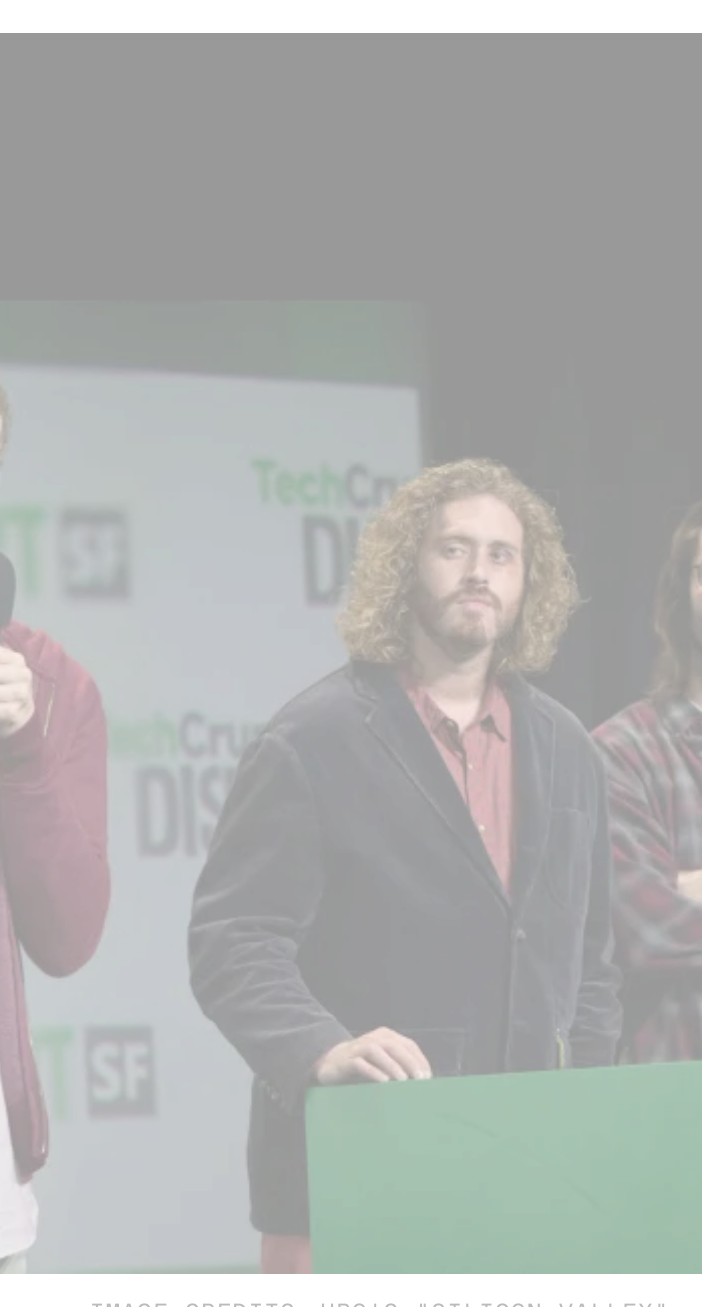

IMAGE CREDITS: HBO'S "SILICON VALLEY"

AI

Google unveils TurboQuant, a lossless AI memory compression algorithm — and yes, the internet is calling it 'Pied Piper'

Sarah Perez 1:38 PM PDT · March 25, 2026

If Google's AI researchers had a sense of humor, they would have called TurboQuant, the new, ultra-efficient AI memory compression algorithm announced Tuesday, "Pied Piper" — or, at least that's what the internet thinks.

The joke is a reference to the fictional startup Pied Piper that was the focus of HBO's "Silicon Valley" TV series that ran from 2014 to 2019.

The show followed the startup's founders as they navigated the tech ecosystem, facing challenges like competition from larger companies, fundraising, technology and product issues, and even (much to our delight) wowing the judges at a fictional version of TechCrunch Disrupt.

Pied Piper's breakthrough technology on the TV show was a compression algorithm that greatly reduced file sizes with near-lossless compression. Google Research's new TurboQuant, is also about extreme compression without quality loss, but applied to a core bottleneck in AI systems. Hence, the comparisons.

> KALEO @CryptoKaleo · Follow
>
> So Google TurboQuant is basically Pied Piper and just hit a Weismann Score of 5.2
>
> Watch on X
>
> Google Research @GoogleResearch
> Introducing TurboQuant: Our new compression algorithm that reduces LLM key-value cache memory by at least 6x and delivers up to 8x speedup, all with zero accuracy loss, redefining AI efficiency. Read the blog to learn how it achieves these results: goo.gle/4bsq2ql
>
> 3:47 PM · Mar 25, 2026
>
> 1.6K Reply Copy link
>
> Read 32 replies

Most Popular

Someone has publicly leaked an exploit kit that can hack millions of iPhones

Cursor admits its new coding model was built on top of Moonshot AI's Kimi

Delve accused of misleading customers with 'fake compliance'

An exclusive tour of Amazon's Trainium lab, the chip that's won over Anthropic, OpenAI, even Apple

Cyberattack on vehicle breathalyzer company leaves drivers stranded across the US

Jeff Bezos reportedly wants $100 billion to buy and transform old manufacturing firms with AI

Employees had to restrain a dancing humanoid robot after it went wild at a California restaurant

Google Research described the technology as a novel way to shrink AI's working memory without impacting performance. The compression method, which uses a form of vector quantization to clear cache bottlenecks in AI processing, would essentially allow AI to remember more information while taking up less space and maintaining accuracy, according to the researchers.

They plan to present their findings at the ICLR 2026 conference next month, along with the two methods that are making this compression possible: the quantization method PolarQuant and a training and optimization method called QJL.

> Justin Trimble @justintrimble · Follow
>
> TurboQuant is the new Pied Piper 🤣
>
> Watch on X
>
> GIF
>
> 6:06 PM · Mar 25, 2
>
> 4 Reply

> Shivang @whyshiva
>
> So basically Tu
>
> Google Research @GoogleResearch
> Introducing TurboQuant: Our new compression algorithm that reduces LLM key-value cache memory by at least 6x and delivers up to 8x speedup, all with zero accuracy loss, redefining AI efficiency. Read the blog to learn how it achieves these results: goo.gle/4bsq2ql
>
> 7:34 PM · Mar 25, 2026
>
> 1 Reply Copy link
>
> Read more on X

Understanding the math involved here is something researchers and computer scientists may be able to do, but the results are exciting the wider tech industry as a whole.

If successfully implemented in the real world, TurboQuant could make AI cheaper to run by reducing its runtime "working memory" — known as the KV cache — by "at least 6x."

Some, like Cloudflare CEO Matthew Prince, are even calling this Google's DeepSeek moment — a reference to the efficiency gains driven by the Chinese AI model, which was trained at a fraction of the cost of its rivals on worse chips, while remaining competitive on its results.

> Matthew Prince @eastdakota · Follow
>
> This is Google's DeepSeek. So much more room to optimize AI inference for speed, memory usage, power consumption, and multi-tenant utilization. Lots of teams at @Cloudflare focused on these areas. #staytuned
>
> Google Research @GoogleResearch
> Introducing TurboQuant: Our new compression algorithm that reduces LLM key-value cache memory by at least 6x and delivers up to 8x speedup, all with zero accuracy loss, redefining AI efficiency. Read the blog to learn how it achieves these results: goo.gle/4bsq2ql
>
> 4:26 PM · Mar 25, 2026
>
> 265 Reply Copy link
>
> Read 9 replies

> Monali @monali_dambre · Follow
>
> Well, we all know who stole the Pied Piper codebase now
>
> Google Research @GoogleResearch
> Introducing TurboQuant: Our new compression algorithm that reduces LLM key-value cache memory by at least 6x and delivers up to 8x speedup, all with zero accuracy loss, redefining AI efficiency. Read the blog to learn how it achieves these results: goo.gle/4bsq2ql
>
> 6:47 PM · Mar 25, 2026
>
> 2 Reply Copy link
>
> Read more on X

Still, it's worth noting that TurboQuant hasn't yet been deployed broadly; it's still a lab breakthrough at this time.

That makes comparisons with something like DeepSeek, or even the fictional Pied Piper, more difficult. On TV, Pied Piper's technology was going to radically change the rules of computing. TurboQuant, meanwhile, could lead to efficiency gains and systems that require less memory during inference. But it wouldn't necessarily solve the wider RAM shortages driven by AI, given that it only targets inference memory, not training — the latter of which continues to require massive amounts of RAM.

> Joe @JoeBGrech · Follow
>
> Pied Piper would have been a better name
>
> Google Research @GoogleResearch
> Introducing TurboQuant: Our new compression algorithm that reduces LLM key-value cache memory by at least 6x and delivers up to 8x speedup, all with zero accuracy loss, redefining AI efficiency. Read the blog to learn how it achieves these results: goo.gle/4bsq2ql
>
> 6:57 PM · Mar 25, 2026
>
> 2 Reply Copy link
>
> Read 1 reply

Topics: AI AI Google pied piper turboquant

Sarah Perez
Consumer News Editor |

Sarah has worked as a reporter for TechCrunch since August 2011. She joined the company after having previously spent...

View Bio

Loading the next article

Advertisement

© 2026 TechCrunch Media LLC.

TechCrunch

TechCrunch asks for your consent to use your personal data to:

Personalised advertising and content, advertising and content measurement, audience research and services development

Store and/or access information on a device

Your personal data will be processed and information from your device (cookies, unique identifiers, and other device data) may be stored by, accessed by and shared with 212 partners, or used specifically by this site. We and our partners may use precise geolocation data. List of partners.

Some vendors may process your personal data on the basis of legitimate interest, which you can object to by managing your options below. Look for a link at the bottom of this page or in the site menu to manage or withdraw consent in privacy and cookie settings.