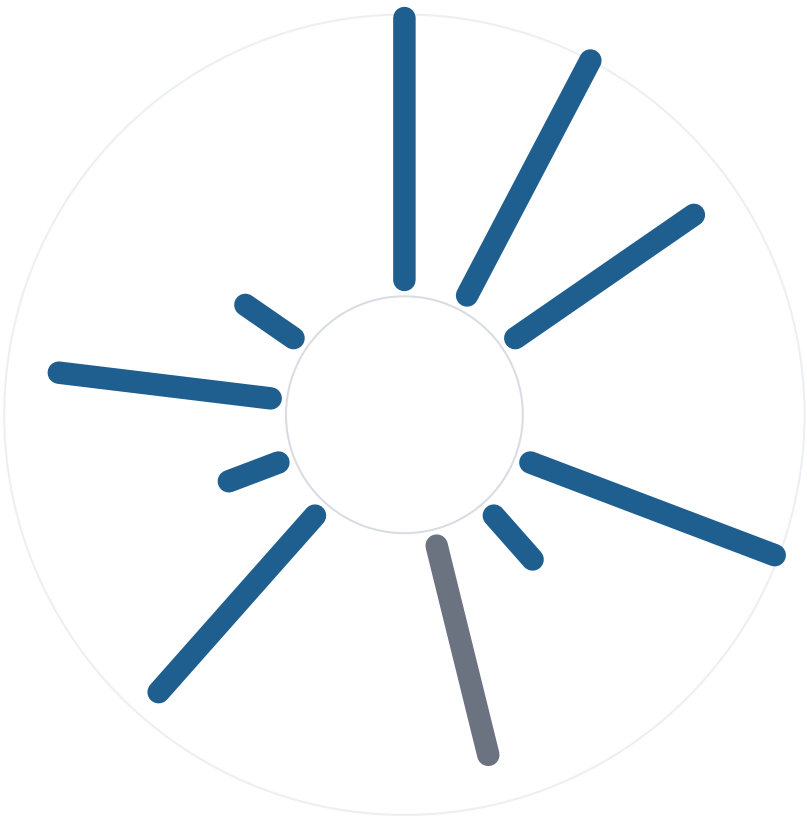
NEW HORIZON · AI DIGEST

The Long Read

28 November 2026

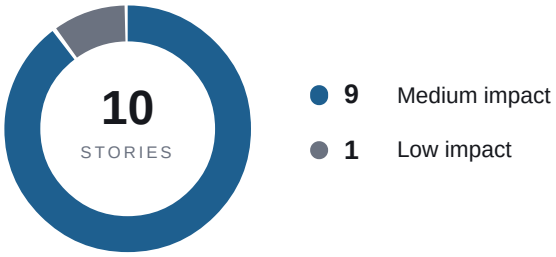
10	3	2599	0
STORIES	SECTIONS	WORDS	HIGH IMPACT

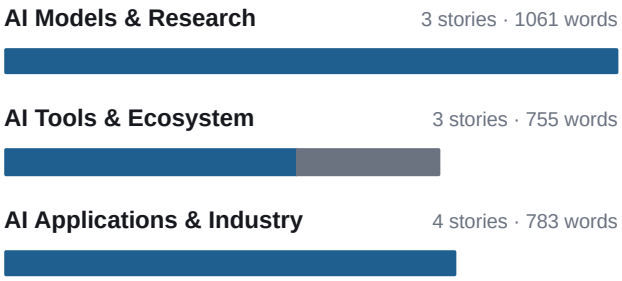


The full-length companion to the daily New Horizon AI Digest.
Every story in today's email, reported at length · new-horizon.tech/digest

The issue at a glance

10 stories · 2599 words · 3 sections · 1 charted





AI Models & Research

01 Alibaba’s voice models now recognise 59 languages

02 JetBrains releases a coding model that skips the deliberation

03 One small AI model corrects itself using 7.8 times as many tokens

AI Tools & Ecosystem

04 Amazon’s Trainium3 chips generate a ten-second film in 23 seconds

05 Claude can now fetch work instructions from connected services

06 Datasette adds interactive chart demos after a Claude-built test

AI Applications & Industry

07 Japan proposes penalties for AI providers that ignore safety orders

08 OpenAI accuses Apple of blocking its device suppliers

09 China’s first widely released AI film flops at the box office

10 Russia unveils a new GigaChat model as sanctions squeeze AI work

How to read this. Every story in today's email is reported here at full length, in the same order. Impact is the writer's judgement of whether a story changes what a reader — in business, at work, as an investor or as a builder — should do or believe this week. Charts appear only where the source itself puts comparable numbers side by side; nothing is estimated to fill a gap. Sources are listed in full on the final page.

SECTION 1 OF 3

AI Models & Research

3 stories ● 3 medium

01 MEDIUM IMPACT thehype.

Alibaba’s voice models now recognise 59 languages

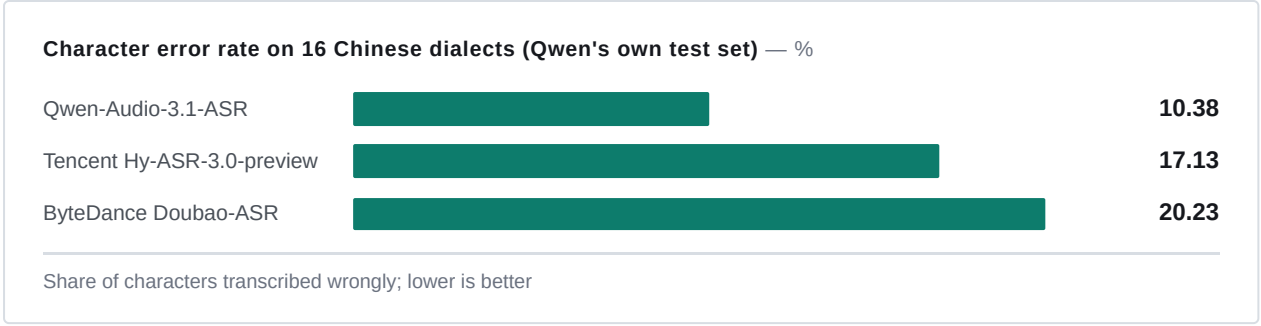
> Alibaba's Qwen team widened its Qwen-Audio-3.1 speech models to cover far more languages, with the recognition model going from 30 languages to 59.

The update, announced November 26, 2026, applies to all three models in the family: one that turns speech into text, one that turns text into speech, and one that holds live spoken conversations. The text-to-speech model now covers 65 languages and the real-time model 53. All three are live through the API on Qwen Cloud, Alibaba's model platform, and none of the weights have been released, so developers rent them rather than run them.

The pitch is Chinese dialects, which most speech models still handle poorly. Qwen's October technical report tested its recognition model on 16 dialects, including Cantonese, Shanghainese, Sichuanese, Hakka and city dialects such as Wenzhou and Suzhou. It got 10.38% of characters wrong on average, against 20.23% for ByteDance's Doubao-ASR and 17.13% for Tencent's Hy-ASR-3.0-preview. The gap is widest where it is hardest: on Wenzhou speech, which is notoriously difficult even for other Chinese speakers, Qwen's error rate was 19.21% against 68.22% for Doubao and 64.74% for Tencent. On public dialect benchmarks the three were close, with Qwen at 4.55%, Tencent at 4.61% and Doubao at 7.15%.

Those are Qwen's own tests, run before this week's update, and no new scores have been published for the added languages. The company also has not listed the languages or named the dialects, and its counting has shifted: the recognition model was previously described as covering 30 languages and 16 dialects, and now lists 8 dialects with no explanation. The October report grouped dialect tests into eight regions, which may explain the lower figure. Buyers should treat the language expansion as unproven until independent tests appear and check it on their own recordings.

Prices are low per million tokens: $0.15 in and $0.47 out for the recognition model, $0.23 in and $1.87 out for text-to-speech, and $0.93 in and $1.87 out for the real-time model. The recognition model also translates dialect speech into standard Mandarin text, labels who is speaking, and can be primed with names or terms in advance. The real-time model listens and speaks at once, handles interruptions and calls tools mid-conversation; in Qwen's September report it completed 82.0% of tasks on an adapted agent benchmark, up from 78.4%, and responded to background chatter in 13% of test cases, down from 73%.



KEY FACTS

RECOGNITION LANGUAGES
59, up from 30

TEXT-TO-SPEECH LANGUAGES
65

REAL-TIME CONVERSATION LANGUAGES
53

AVERAGE CHARACTER ERROR RATE ON 16 CHINESE DIALECTS (QWEN'S OWN TEST)
10.38%, vs 20.23% for ByteDance Doubao and 17.13% for Tencent

WENZHOU DIALECT ERROR RATE
19.21%, vs 68.22% Doubao and 64.74% Tencent

RECOGNITION MODEL PRICE
$0.15 per 1M input tokens, $0.47 per 1M output tokens

WHY IT MATTERS

For anyone building voice products for Chinese users, Qwen now looks like the strongest option on dialects, at token prices low enough to test cheaply. The doubled language count also makes it more usable outside China, but with no published results for the new languages, teams should benchmark it on their own audio first.

Read the original at thehype. →

---

02 MEDIUM IMPACT The JetBrains Blog

JetBrains releases a coding model that skips the deliberation

> JetBrains has released Mellum2.1 Instruct, a version of its open coding model that gives answers without working through them step by step first, aimed at the routine chores that fill much of a coding agent's day.

The company released Mellum2.1 Thinking in October, trained to reason its way through problems as a coding agent. The new Instruct variant shares the same design and training but skips the reasoning, producing straight answers. JetBrains says that on agentic coding tasks it generates about half as many tokens as the Thinking version while performing almost as well on SWE-bench Verified, a standard test of whether a model can fix real software bugs. Tokens are the units of text a model produces, and each one costs time and computing.

The pitch is speed and cost, especially for anyone running the model on their own hardware. JetBrains says it received many requests from the local-inference community for a more concise model for moderately hard tasks after the Thinking release. Much of a coding agent's work is small: a minor fix, a command to run, a question about the code. Reasoning in those cases adds time and computing without adding quality, the company says. For hard algorithmic and maths problems, it concedes, Thinking remains the better choice.

Against its predecessor, Mellum2 Instruct, the biggest gain is on SWE-bench Verified, where the score doubled. The new training also improved code generation, where JetBrains says it now leads a comparison group that includes Qwen3.5 9B and Gemma 4 E4B, both run with thinking turned off, and multi-step tool use, where it is level with Qwen3.5 9B. Qwen3.5 9B still scores higher on SWE-bench Verified, but JetBrains says Mellum2.1 generates tokens about 1.4 times faster for a single request and about 1.8 times faster under load, because it activates only 2.5 billion of its 12 billion parameters per token.

The most striking detail is the training setup. JetBrains says a single run collected about half a million agent trajectories from 10,000 concurrent sandboxes, using NVIDIA's open-source NeMo training libraries across clusters of H200 and B300 GPUs. Both variants are on Hugging Face under the Apache 2.0 licence, with versions packaged for running on your own machine and a multi-token prediction add-on that JetBrains says makes generation up to 1.8 times faster. The release is an incremental addition to an existing family rather than a new class of model, but it is a free, openly licensed option for teams building coding tools they host themselves.

KEY FACTS

RELEASED
November 2026, following Mellum2.1 Thinking in October

LICENCE
Apache 2.0, available on Hugging Face

TOKEN USE
About half as many as the Thinking variant on agentic coding tasks

SPEED
Up to 1.8x faster generation with the multi-token prediction head

TRAINING SCALE
About 500,000 agent trajectories from 10,000 concurrent sandboxes in one run

BENCHMARK GAIN
SWE-bench Verified score doubled versus Mellum2 Instruct

> **WHY IT MATTERS**
>
> Teams building coding assistants can cut the time and computing they pay for routine requests by using a model that answers directly, while keeping a reasoning model for the hard cases. Because it is openly licensed and small enough to run locally, it also gives companies an option that does not depend on a vendor's servers.

Read the original at The JetBrains Blog →

---

03 MEDIUM IMPACT MachineLearningMastery.com

One small AI model corrects itself using 7.8 times as many tokens

> A small open-source AI model, made to check and correct its own work in three passes, turned a wrong answer into a right one — but spent 7.8 times as many tokens and ran 6.3 times slower than answering in one go.

The demonstration, published as a Python walkthrough by MachineLearningMastery.com, used a single model — Qwen2.5-7B-Instruct, a freely available model small enough to run on a free Google Colab graphics chip — to play three roles in turn. A "Proposer" gave a first answer, a "Critic" worked the problem out independently and compared, and a "Judge" issued the final reply. The same model wore all three hats, with its instructions swapped between steps, because loading three separate models at once would have run out of memory.

The test question was a classic trap: Bob is 20, Alice is half his age; how old is Alice when Bob is 100? The single-pass model said 50, dividing 100 by two and ignoring that the age gap stays fixed. The Critic caught it, and the Judge returned 90. The cost was measured alongside: 85 tokens and 2.63 seconds for the single pass, against 660 tokens and 16.67 seconds for the full loop.

That gap is the story for anyone paying for AI by the token or waiting on a response. Chaining models together — a pattern sold by multi-agent frameworks such as AutoGen and LangGraph — adds what the article calls a "token tax", and the bill compounds with every step added. The author's conclusion is that the loop is wasted on simple factual queries and earns its keep only on problems that need reasoning, where the gain is robustness rather than speed.

The most surprising detail is that the correction came from a prompt instruction, not a bigger model: the Critic is told to calculate the answer itself step by step before judging. The technique is not new — the three-role pattern is established, and the model is off the shelf. What the piece adds is a cheap, reproducible way to put a number on the trade-off before committing to it.

KEY FACTS

TOKEN COST OF THE THREE-STEP LOOP
7.8x a single pass

LATENCY OF THE THREE-STEP LOOP
6.3x slower

TOKENS USED, SINGLE PASS VS LOOP
85 vs 660

TIME TAKEN, SINGLE PASS VS LOOP
2.63s vs 16.67s

MODEL USED
Qwen2.5-7B-Instruct, 4-bit quantized

HARDWARE
Google Colab Nvidia T4 GPU (16GB)

> **WHY IT MATTERS**
>
> Teams building with AI now have a simple way to price self-checking: measure tokens and seconds against a single pass before deciding a multi-step loop is worth it. For most routine queries, the arithmetic in this test says it is not.

Read the original at MachineLearningMastery.com →

SECTION 2 OF 3

AI Tools & Ecosystem

3 stories · 2 medium · 1 low

04 MEDIUM IMPACT huggingface.co

Amazon’s Trainium3 chips generate a ten-second film in 23 seconds

> A developer given a month on an Amazon Trainium3 server used the last five days to generate video, and got a 10-second clip with sound out in 22 to 23 seconds — roughly the same as the paid online service of Lightricks, the company behind the video model.

The test came from Hugging Face, which had access to a Trainium3 machine for a month through its partnership with Amazon Web Services. With five days left, engineer Daniel van Strien (writing on the Hugging Face blog) tried LTX-2.5, a model from Lightricks that generates video and sound together, and built short films from a script rather than single clips. Once the models were loaded, a 10-second clip at 2048×1024 took 22 to 23 seconds on half of the machine. Lightricks says its own online API makes a 10-second 1080p clip in 23.7 seconds.

The first attempt had taken 5 minutes 20 seconds for a lower-resolution clip, with the chips actually computing for only about 45 seconds of that. Five days of work cut the wall time roughly 14-fold. Most of the gain came from fixing waste on the machine hosting the chips rather than the chips themselves: every processor was re-encoding the same text prompt, and the video was being decoded slowly on ordinary CPUs. Moving the decode onto the chips cut it from 357 seconds to 1.6 seconds.

The most surprising result is not speed. The model can take the first second of real camera footage and invent what happens next. Given a clip of a humanoid robot reaching for an orange, it invented the grab on its own, keeping the table, hands and fruit consistent. Changing only the prompt, it produced a person leaning in to carry the orange alongside the robot — five different takes, each reproducible from a saved file and seed. That matters to anyone training robots or self-driving systems, who need video of rare or risky events they cannot easily film. The catch: the model follows prompts, not the robot's commands, so a training setup would need a team to fine-tune it on its own robots' actions.

What is new here is that video generation runs credibly on Amazon's chips at all, not just on Nvidia's. What is incremental is the output itself: shots last at most 20 seconds, each shot makes its own sound with no carry-over, and the model still gets details wrong — the same prompt gave a goat funny horns in three takes out of four. The author calls the results drafts.

KEY FACTS

TIME FOR A 10-SECOND CLIP WITH SOUND
22-23 seconds

LIGHTRICKS' CLAIMED API TIME FOR A 10-SECOND 1080P CLIP
23.7 seconds

FIRST ATTEMPT, LOWER RESOLUTION
5 minutes 20 seconds

SPEED-UP AFTER FIVE DAYS OF WORK
14x

CHIPS USED FOR ONE CLIP
32 of 64

LONGEST SINGLE SHOT
20 seconds

> WHY IT MATTERS
>
> Video generation has been effectively tied to Nvidia hardware; a credible run on Amazon's chips gives buyers a second supplier and a price lever. For studios and robotics teams, the practical draw is reproducible drafts — several takes of a shot from one saved file — not finished footage.

Read the original at huggingface.co →

---

05 MEDIUM IMPACT venturebeat.com

Claude can now fetch work instructions from connected services

> Claude can now pull work instructions from connected services, not just the tools those services expose.

Source not retrievable. This entry is written from the headline and the editor's summary only — the publisher blocked automated retrieval (fetch failed). Follow the link for the full report.

The capability arrives as a public beta under the Model Context Protocol, the standard for plugging outside services into AI assistants. Servers can now hand over workflow guides alongside their tools; Shopify, HubSpot and Atlassian are among the examples given. Because those instructions can change after a company has approved them, the standard requires fresh approval whenever a previously accepted skill is altered. That shifts what businesses are signing off on when they connect a server.

> WHY IT MATTERS
>
> If approved instructions can be rewritten later, connecting a server becomes an ongoing review commitment rather than a one-time decision, which may slow adoption inside cautious organisations.

Read the original at venturebeat.com →

---

06 LOW IMPACT Simon Willison's Weblog

Datasette adds interactive chart demos after a Claude-built test

> The developer of Datasette, an open-source tool for publishing data, has released a plugin for interactive charts and rebuilt its design after an AI-written test plugin exposed where it was awkward to use.

Simon Willison released the datasette-charts plugin with a revised extension hook — the mechanism that lets other plugins add their own chart types. He had not used the hook for anything before asking Claude Opus 5.5, an AI model, to build an experimental plugin against it. That exercise surfaced "a bunch of friction points in the design," and the release addresses them so that building additional chart plugins should be "a whole lot more ergonomic."

The audience is narrow but specific: people who publish datasets with Datasette and want readers to explore the numbers rather than download a file. For them the change is practical rather than dramatic — a plugin author who previously had to work around the hook's rough edges should find the job easier. The wider readership of business users looking at a published dataset sees the difference only indirectly, in whether a chart type they want exists at all.

The most telling detail is the method. The hook was tested by having an AI model write a plugin against it, and the failures that turned up became the specification for the redesign. That is a cheap way to find out where an interface confuses its users: hand it to something that has no context and see where it stumbles. It says nothing about whether the resulting charts are better than a rival's, only that the extension point was reworked.

Willison also launched a site for the plugin at charts.datasette.io, with a tutorial for plugin authors and a gallery of the core charts and annotations. Each item in the gallery links to an interactive demo running on Datasette Lite, a browser-based version of the tool. The genuinely new parts are the revised hook and the site; the charts themselves are an incremental step for an existing open-source project, not a new product category.

KEY FACTS

RELEASE DATE
27th November 2026

MODEL USED FOR THE TEST PLUGIN
Claude Opus 5.5

NEW WEBSITE
charts.datasette.io

SITE CONTENTS
Tutorial for plugin authors and a gallery of core charts and annotations

DEMOS RUN ON
Datasette Lite, in the browser

> **WHY IT MATTERS**
>
> If you publish data with Datasette, chart plugins should be easier to build and there is now a tutorial and working demos to copy from. For everyone else it is a small open-source release with no bearing on budgets or buying decisions.

Read the original at Simon Willison’s Weblog →

SECTION 3 OF 3

AI Applications & Industry

4 stories · 4 medium

07 MEDIUM IMPACT The Japan Times

Japan proposes penalties for AI providers that ignore safety orders

> A Japanese government AI strategy panel has outlined a draft bill that would give the state power to issue binding orders to AI service providers and penalise those that ignore them.

The outline, reported by The Japan Times on Friday, would amend the AI law enacted in May 2025. That law was written mainly to encourage use of AI; it lets the government investigate AI service providers and offer guidance, but its authority is described as weak. The draft adds teeth: operators would have to report serious AI-related incidents to the government, and the state could make recommendations, issue orders, and punish businesses that do not comply.

The draft also obliges providers of advanced AI services to take measures against cyberattacks. The Japan Times reports this is apparently aimed at OpenAI, which offers the ChatGPT chatbot, and Anthropic, developer of the Claude chatbot — the two best-known makers of generative AI, the kind that produces text on request. The source does not say what specific security measures would be required, what the penalties would be, or when they would apply. The government aims to finalise the bill early on the basis of the outline and submit it to parliament.

The backdrop is a rise in cyberattacks that exploit fast-moving AI technology, according to the source. That framing puts Japan alongside other governments trying to move from encouraging AI to policing it, though the source gives no comparison with rules elsewhere.

What is genuinely new is the shift from guidance to enforceable orders, and the incident-reporting duty, which would put AI providers on the same footing as firms in other regulated industries that must tell regulators when something goes wrong. What is incremental is the law itself: this amends a statute less than a year old rather than starting fresh.

For companies selling AI services in Japan, the practical question is how broadly "advanced AI services" is defined and how heavy the penalties are — neither is settled in the outline. For investors, the signal is that Japan intends to treat large AI providers as regulated operators, not just technology vendors. The bill still has to be finalised and passed by parliament before any of it takes effect.

KEY FACTS

LAW ENACTED
May 2025

NEW POWERS
Recommendations, orders and penalties for non-compliance

NEW DUTY
Report serious AI-related incidents to government

SECURITY REQUIREMENT
Advanced AI providers must take measures against cyberattacks

NAMED IN DRAFTING
OpenAI (ChatGPT) and Anthropic (Claude)

NEXT STEP
Bill to be finalised and submitted to parliament

> WHY IT MATTERS
>
> Companies offering AI services in Japan could soon face binding government orders on security and mandatory incident reporting, with penalties for non-compliance — a compliance cost that did not exist under the 2025 law.

Read the original at The Japan Times →

---

08 MEDIUM IMPACT www.lemonde.fr

OpenAI accuses Apple of blocking its device suppliers

> OpenAI has countersued Apple, accusing it of pressuring shared manufacturers and obtaining confidential plans for OpenAI's unreleased devices through a former employee.

Source not retrievable. This entry is written from the headline and the editor's summary only — the publisher blocked automated retrieval (extracted only 46 words (paywall/consent wall?)). Follow the link for the full report.

The countersuit, reported by Le Monde, claims Apple leaned on manufacturers that work with both companies and got hold of confidential plans for OpenAI devices that have not been released. Apple's response: the filing is a distraction from its own lawsuit, which accuses former staff of stealing trade secrets. The two cases now run in parallel, and the suppliers named in neither account are the ones most exposed.

> WHY IT MATTERS
>
> It pulls the hardware supply chain into a dispute between two companies that both depend on outside manufacturers, and it signals that OpenAI's unreleased devices are far enough along to be worth fighting over in court.

Read the original at www.lemonde.fr →

---

09 MEDIUM IMPACT The Hollywood Reporter

China’s first widely released AI film flops at the box office

> The first feature-length AI-generated film to get a wide theatrical release in a major market, China's sci-fi thriller Sanxingdui: Future Past, opened to an estimated $198,000, about 12 percent of its

roughly $1.5 million budget.

The film, a 100-minute sci-fi thriller inspired by archaeological finds from the Sanxingdui civilization, went out last week with more than 10,000 screenings on its opening day, according to the French wire service Agence France-Presse. That scale is what makes the result notable: earlier AI-made features, such as India's 95-minute musical Love You in May 2025 and Malaysia's Pirate Queen: Zheng Yi Sao in January, played in limited release. Future Past is the first longform AI film to get a true wide release with reported box office numbers in a major territory, and so the first real commercial test of the technology for feature-length projects.

Jiang Defu, chief executive of the production company Bona Film Group, told the Chinese film outlet Yuli that the film was still an important experiment. "We are the first to try it, the first to get a theatrical release permit and the first to be tested by the audience," he said. "I think that is not easy at all." One moviegoer was blunter: "In terms of quality, compared with human production, it still falls short."

The poor opening lands in the same week that the first chapter of Elon Musk's AI-produced version of The Odyssey was turning off viewers in the United States. The next test comes there on Dec. 4, when Gods Don't Give Gifts, an R-rated anthology of four sci-fi stories created by YouTuber Gossip Goblin and distributed by Fable, is scheduled for limited release.

What is new here is the distribution, not the filmmaking: AI has already taken over China's short online films, known as microdramas, 95 percent of which are AI-generated, according to Fortune. A feature-length AI hit looks like a matter of time rather than a question of whether the tools work. For now, the audience has priced the experiment at about $198,000.

KEY FACTS

OPENING BOX OFFICE
$198,000 (estimated)

BUDGET
10 million yuan (about $1.5 million)

SHARE OF BUDGET RECOUPED
About 12 percent

OPENING-DAY SCREENINGS
More than 10,000

RUNTIME
100 minutes

CHINA'S AI-GENERATED MICRODRAMAS
95 percent

WHY IT MATTERS

Studios and investors now have a first hard number on what AI-made features earn at the box office, and it is small: cost savings on production do not create an audience. The near-term money in AI video is in cheap short-form content, not cinema.

Read the original at The Hollywood Reporter →

---

10 MEDIUM IMPACT nytimes.com

Russia unveils a new GigaChat model as sanctions squeeze AI work

Sberbank has unveiled a new version of its GigaChat model at Russia's AI Journey conference, claiming it comes close to leading Chinese models.

Source not retrievable. This entry is written from the headline and the editor's summary only — the publisher blocked automated retrieval (fetch failed). Follow the link for the full report.

The presentation is a statement of intent as much as a product launch. Sberbank says the latest GigaChat approaches the leading Chinese models, a claim this desk could not check. The conference also aired the obstacles: American and Chinese hardware costs more, skilled staff are leaving, and Ukrainian drone strikes have hit data centres. For business readers, the takeaway is that Russia's AI effort continues, but under conditions that make catching the frontier expensive and slow.

> **WHY IT MATTERS**
>
> If Sberbank's claim holds, Russian firms get a domestic chatbot option that sanctions have made harder to source elsewhere; if it does not, Russian users stay dependent on foreign models they may not be able to buy.

Read the original at nytimes.com →

Sources

01 Alibaba’s voice models now recognise 59 languages
https://thehype.news/qwen-audio-3-1

02 JetBrains releases a coding model that skips the deliberation
https://blog.jetbrains.com/ai/2026/11/mellum2-1-instruct-is-out-bringing-faster-answers-for-coding-agents

03 One small AI model corrects itself using 7.8 times as many tokens
https://machinelearningmastery.com/building-a-multi-agent-verification-loop-in-python-with-small-open-models/

04 Amazon’s Trainium3 chips generate a ten-second film in 23 seconds
https://huggingface.co/blog/video-trainium3

05 Claude can now fetch work instructions from connected services
https://venturebeat.com/technology/mcp-servers-can-now-deliver-ai-agent-skills-changing-what-companies-approve-when-they-connect-one

06 Datasette adds interactive chart demos after a Claude-built test
https://simonwillison.net/2026/Nov/27/datasette-charts/

07 Japan proposes penalties for AI providers that ignore safety orders
https://www.japantimes.co.jp/news/2026/11/28/japan/politics/japan-revise-ai-law-penalties

08 OpenAI accuses Apple of blocking its device suppliers
https://www.lemonde.fr/en/pixels/article/2026/11/28/openai-countersues-apple-in-dispute-over-trade-secret-thefts_6790291_13.html

09 China’s first widely released AI film flops at the box office
https://www.hollywoodreporter.com/movies/movie-news/sanxingdui-future-past-ai-movie-box-office-flop-1236742696

10 Russia unveils a new GigaChat model as sanctions squeeze AI work
https://www.nytimes.com/2026/11/27/world/europe/russia-ai-journey-conference.html

---

About this document. Every story in the 28 November 2026 New Horizon AI Digest, reported at length. Each entry is written from the publisher's own article text; where a source could not be retrieved the entry is explicitly marked and kept short rather than padded.

Images and licensing. Figures are used only where the source licence permits redistribution, and are credited in the caption. Publisher artwork is not reproduced. All titles link to the original publication.

Vocabulary. Common AI terms are defined at new-horizon.tech/glossary/.

Generated 2026-11-28T06:36:41Z · model deepseek-v4.1-flash:cloud · New Horizon, new-horizon.tech · To unsubscribe or manage your subscription, see the footer of the daily email.