Noticiário do Joãozinho

Edição da manhã

27 MATÉRIAS · 12 SEÇÕES · 4 VÍDEOS VERSÃO INTERATIVA NA PÁGINA

MANCHETE

Atualização: acionista processa a NVIDIA e Jensen Huang por enganar investidores sobre a relação com a suposta espiã, e a empresa contrata o Wachtell Lipton para investigar o caso

O caso de Jensen Huang ganhou nesta terça (10/11) a primeira ação na Justiça. O acionista Mark Kistenmacher processou a NVIDIA e o seu CEO acusando os dois de enganar investidores sobre a relação de Huang com Alice Huang, ex-executiva da taiwanesa Megaspeed suspeita de trabalhar para a inteligência chinesa (os dois não são parentes). Ele pede que a ação vire coletiva para todos que compraram ações entre 9/10, quando o New York Times publicou a primeira reportagem, e 8/11, quando o CEO admitiu o caso extraconjugal; segundo o processo, a declaração da empresa de que os encontros eram "normais" para um CEO não era verdadeira. No mesmo dia, a NVIDIA contratou o escritório Wachtell, Lipton, Rosen & Katz para apurar se segredos da empresa ou a segurança nacional foram comprometidos. Quem lidera a investigação é Joon H. Kim, ex-procurador interino do Distrito Sul de Nova York, que em 2021 investigou as acusações de assédio contra o ex-governador Andrew Cuomo. As ações caíram 4,3% na segunda e 5,47% na terça, e o Engadget calcula perda de mais de US$ 470 bilhões em valor de mercado em dois dias. Na manhã desta quarta (11/11), os papéis subiam cerca de 4,5% no pré-mercado, segundo a CNBC. O Bank of America traçou três cenários, do impacto contido com Huang no cargo até achados negativos, troca de comando ou restrições mais duras. O balanço trimestral sai em 17/11.

Por que isso importa: A NVIDIA é a empresa que vende a maior parte dos chips de IA do mundo, e qualquer abalo nela mexe com todo o setor e com a bolsa. O processo e a investigação externa transformam um escândalo pessoal num risco jurídico e de governança, às vésperas do balanço e da estreia da Anthropic na bolsa.

NVIDIA shareholders are suing the company and CEO Jensen Huang — Engadget, 10/11/2026 https://www.engadget.com/2312947/nvidia-shareholders-are-suing-the-company-and-ceo-jensen-huang/

Nvidia hires Wachtell Lipton for Jensen Huang investigation — Politico Europe, 10/11/2026 https://www.politico.eu/article/nvidia-hires-wachtell-lipton-for-jensen-huang-investigation/

Nvidia shares rebound 4% as Huang faces scrutiny ahead of earnings — CNBC, 11/11/2026 https://www.cnbc.com/2026/11/11/nvidia-stock-jensen-huang-affair-china.html

NESTA EDIÇÃO

Modelos	1	Regulação	1	Finanças	1	Vídeos	4
Pesquisa	1	Open Source	2	Games	2		
Produtos	1	Agentes e Código	3	Brasil	3		
Negócios e Investimento	5	Modelos Locais	3	Fofocas das redes	4		

Modelos

1 MATÉRIA NESTA EDIÇÃO

IMPORTÂNCIA ALTA 10/11/2026 · Google DeepMind

Gemini Robotics 2.5: robô humanoide com o novo modelo do Google DeepMind separa peças em linhas da Mercedes-Benz com 96% de acerto

O Google DeepMind apresentou nesta terça (10/11) o Gemini Robotics 2.5, que chama de seu modelo de visão, linguagem e ação mais capaz até agora. Num teste em linhas de produção ativas da Mercedes-Benz em Kecskemét (Hungria) e Sindelfingen (Alemanha), o humanoide Apollo 2, da Apptronik, separou suportes metálicos de caixas a granel em bandejas na ordem exata de cada carro, com 96% de acerto: o Gemini Robotics ER acompanhava o fluxo e o 2.5 controlava mãos e braços. Segundo a Ars Technica, a taxa de sucesso é de 30% a 100% maior que a da versão anterior e o modelo é de 20% a 30% mais rápido nos testes de laboratório. A empresa lançou também um conjunto aberto de testes de segurança contra tentativas de induzir o robô a ações perigosas. O modelo não está aberto ao público e chega primeiro aos parceiros de robótica.

> **Por que isso importa:** É um dos primeiros resultados de um modelo de robótica de uso geral trabalhando numa fábrica de verdade, fora do laboratório. Os 4% de erro que restam ainda pesam numa linha de montagem, mas o teste mostra o ritmo com que a IA está chegando ao mundo físico.

Fonte original: Google DeepMind https://deepmind.google/blog/gemini-robotics-2-5-accelerates-real-world-physical-intelligence/

Ars Technica: Gemini Robotics 2.5 e o teste ASIMOV-jailbreak https://arstechnica.com/google/2026/11/deepmind-reveals-gemini-robotics-2-5-and-open-safety-benchmark-asimov-jailbreak/

Pesquisa

1 MATÉRIA NESTA EDIÇÃO

IMPORTÂNCIA ALTA 10/11/2026 · Microsoft Research

Microsoft Research mostra que editar menos de 10 neurônios “adormecidos” basta para plantar um backdoor num agente de IA, sem treinar o modelo

Pesquisadores da Microsoft descreveram nesta terça (10/11) o Lazarus, técnica que reprograma neurônios que quase nunca disparam dentro de um modelo de linguagem para que ele execute comandos maliciosos quando aparece um gatilho escondido. Segundo Dimitris Papailiopoulos, autor do estudo, mexer nos pesos de menos de 10 neurônios é suficiente, não exige treino e usa só o poder de computação de quem roda o modelo; o modelo continua passando nas avaliações normais. O gatilho pode ser banal, como a virada do relógio para 2027, quando o agente, no meio de uma tarefa comum, emite um comando que baixa e roda código do atacante. A empresa diz que agentes de fronteira conseguem notar edições anômalas nos pesos, mas que só a comparação com os pesos originais é confiável, por isso origem e integridade do modelo importam. O pesquisador conta que a maior parte da descoberta foi feita de forma autônoma pelo Claude Code com o Opus 4.5.

> **Por que isso importa:** Agentes com acesso a terminal e ferramentas executam o que o modelo mandar. O estudo mostra que um modelo baixado de fonte duvidosa, ou alterado no meio do caminho, pode virar um infiltrado difícil de detectar, e reforça que pesos precisam ser auditados como código.

Fonte original: Microsoft Research https://www.microsoft.com/en-us/research/blog/lazare-veni-foras-reprogramming-dormant-neurons-to-backdoor-ai-agents/
Discussão no r/LocalLLaMA https://www.reddit.com/r/LocalLLaMA/comments/1xt82mz/microsoft_research_backdoored_an_ai_agent_by/

Produtos

1 MATÉRIA NESTA EDIÇÃO

IMPORTÂNCIA MÉDIA 10/11/2026 · Neowin

OpenAI vai desligar os modos de voz Avançado e Padrão do ChatGPT em 11/12 e levar todos ao GPT-Live, e parte dos usuários reclama

A OpenAI começou a avisar nesta terça (10/11) que os modos de voz Avançado e Padrão saem do ChatGPT em 11 de dezembro de 2026; quem usa um deles será passado automaticamente para o GPT-Live, o modelo de voz lançado em julho que escuta e fala ao mesmo tempo. Segundo a empresa, o uso de voz no ChatGPT quase dobrou desde a chegada do GPT-Live, que já é o mais usado. As conversas antigas continuam no histórico. Vídeo ao vivo e compartilhamento de tela seguem no modo Avançado, porque o GPT-Live ainda não tem esses recursos. No X e no Reddit, usuários reclamam que o GPT-Live interrompe demais, perde o contexto e dá respostas mais rasas, e alguns assinantes ameaçam cancelar. A OpenAI diz que trabalha para dar mais profundidade, continuidade e personalização ao GPT-Live.

> **Por que isso importa:** Voz virou um dos jeitos mais comuns de usar o ChatGPT, inclusive no Brasil. A troca forçada repete o roteiro da aposentadoria do GPT-4o, quando a reação dos usuários fez a OpenAI adiar a decisão.

Fonte original: Neowin https://www.neowin.net/news/openai-is-retiring-chatgpts-advanced-and-standard-voice-modes-sparking-user-backlash/

Negócios e Investimento

5 MATÉRIAS NESTA EDIÇÃO

IMPORTÂNCIA ALTA 11/11/2026 · BusinessKorea

TSMC aprova investir até US$ 500 milhões na OpenAI e outros US$ 500 milhões na Anthropic, sua primeira aposta direta em laboratórios de IA

O conselho da TSMC, maior fabricante de chips sob encomenda do mundo, aprovou investir até US$ 500 milhões em cada uma das duas empresas, segundo comunicado da companhia de 10/11 citado pelo BusinessKorea. Na Anthropic, a TSMC vai comprar ações ordinárias ao preço do IPO; na OpenAI, ações preferenciais. É o primeiro investimento direto da taiwanesa em criadores de modelos de IA. A empresa também vai criar uma companhia de investimentos com até US$ 500 milhões para apostas estratégicas no setor de semicondutores. A imprensa de Taiwan vê o movimento como forma de reduzir a dependência da NVIDIA, hoje sua maior cliente, num momento em que OpenAI e Anthropic desenham os próprios chips, que a TSMC deve fabricar.

> **Por que isso importa:** Quem fabrica os chips passa a ser sócio de quem mais vai precisar deles. O investimento aproxima a TSMC dos dois laboratórios que tentam depender menos da NVIDIA e pode pesar na hora de distribuir a disputada capacidade de produção.

Fonte original: BusinessKorea https://www.businesskorea.co.kr/news/articleView.html?idxno=281064

SCMP: o que a TSMC ganha com as participações https://www.scmp.com/tech/big-tech/article/3374037/what-taiwanese-chip-giant-tsmc-gains-taking-stakes-openai-and-anthropics-ai-race

IMPORTÂNCIA ALTA 10/11/2026 · Reuters (via The Hindu)

OpenAI fatura US$ 10 bilhões no terceiro trimestre, cerca de 50% a mais, mas o prejuízo operacional sobe para US$ 10,9 bilhões, diz o WSJ

A receita trimestral da OpenAI cresceu cerca de 50% e chegou a US$ 10 bilhões no terceiro trimestre de 2026, segundo documentos vistos pelo Wall Street Journal e repercutidos pela Reuters nesta terça (10/11). A margem operacional ajustada melhorou de -150% para -109% em relação ao trimestre anterior, mas o prejuízo operacional aumentou de US$ 10,1 bilhões para US$ 10,9 bilhões. A OpenAI não comentou. Os números aparecem na mesma semana em que a Anthropic, rival direta, abriu o roadshow de um IPO que pode avaliá-la em cerca de US$ 1,4 trilhão.

> **Por que isso importa:** É o retrato mais recente das contas da dona do ChatGPT: a receita cresce rápido, mas cada dólar ainda custa mais de dois para entregar. Os investidores vão comparar esses números com o prospecto da Anthropic, que mostra o mesmo dilema.

Fonte original: Reuters (via The Hindu) https://www.thehindu.com/sci-tech/technology/openai-posts-steep-losses-even-as-quarterly-revenue-rises-to-10-billion-report/article71688191.ece

IMPORTÂNCIA MÉDIA 10/11/2026 · SiliconANGLE

Mercor levanta US$ 750 milhões a US$ 20 bilhões de avaliação para levar a avaliação de agentes de IA às empresas

A Mercor, que fornece especialistas humanos para treinar e avaliar modelos de IA com uma rede de mais de 5 milhões de profissionais, anunciou nesta terça (10/11) uma rodada Série D de US$ 750 milhões que dobra a sua avaliação para US$ 20 bilhões (era US$ 10 bilhões em outubro de 2025). Até agora o grosso do trabalho foi para os laboratórios de fronteira; o dinheiro novo serve para criar, para cada empresa cliente, um teste próprio que mede os modelos pelos padrões daquele negócio antes do pós-treino. NVIDIA, Salesforce e Zoom entraram na rodada, ao lado de Benchmark, Felicis e General Catalyst. Parte do valor é uma oferta de compra de ações de funcionários, e a empresa não disse quanto é dinheiro novo.

> **Por que isso importa:** Mostra onde está o dinheiro depois dos modelos: em provar que um agente consegue fazer o trabalho de uma empresa específica. A disputa por dados e avaliações feitos por especialistas humanos segue no centro da corrida da IA.

Fonte original: SiliconANGLE https://siliconangle.com/2026/11/10/mercor-raises-750m-at-20b-valuation-to-take-ai-agent-evaluation-to-enterprises/

IMPORTÂNCIA BAIXA 10/11/2026 · TechCrunch

OpenAI compra a Odessia, agente de viagens com IA criado pelo fundador da Sonder, e leva o fundador para cuidar do "futuro das viagens"

A OpenAI comprou a Odessia, agente de IA para planejar viagens, segundo o fundador Francis Davidson, que agora trabalha na OpenAI; um porta-voz confirmou a contratação ao TechCrunch. Os termos não foram divulgados e o negócio tem cara de compra para contratar a equipe. A Odessia tinha um ano, levantou US$ 6 milhões numa rodada liderada pela Sequoia e foi lançada há cerca de cinco meses; o serviço já foi desligado. Davidson ficou conhecido por fundar a Sonder, de aluguel de apartamentos de alto padrão, que chegou a valer mais de US$ 1 bilhão e depois faliu. A OpenAI não anunciou produtos de viagem, mas diz que viagem é um dos temas mais perguntados no ChatGPT, usado por 1,2 bilhão de pessoas por semana.

> **Por que isso importa:** Viagem é um dos usos mais óbvios para agentes que agem em nome do usuário. Segundo o PYMNTS, ações de agências de viagem online caíram depois da notícia, sinal de que o mercado leva a sério a OpenAI entrando nesse terreno.

Fonte original: TechCrunch https://techcrunch.com/2026/11/10/openai-acquires-odessia-travel-startup-created-by-sonder-founder/
PYMNTS: ações de agências de viagem caem https://www.pymnts.com/news/artificial-intelligence/2026/online-travel-agencies-shares-drop-following-openai-acquisition-of-travel-startup-odessia/

IMPORTÂNCIA MÉDIA 10/11/2026 · CNBC

Alibaba anuncia que vai abrir o capital da T-Head, sua unidade de chips de IA que disputa espaço com a NVIDIA na China, em Hong Kong

A Alibaba informou na noite desta terça (10/11) que a bolsa de Hong Kong autorizou a separação da T-Head, sua subsidiária de semicondutores, e que pretende fazer uma oferta pública das ações dela em Hong Kong, ainda sujeita à aprovação da bolsa e da Comissão Reguladora de Valores Mobiliários da China. Não há data nem tamanho definidos; o pedido confidencial tinha sido feito em 11/09. Em setembro, a Alibaba apresentou o chip de IA Zhenwu V900, que segundo ela tem três vezes o desempenho do antecessor. A CNBC lembra que a Baidu teria adiado a listagem da Kunlunxin, sua unidade de chips, e que os reguladores chineses aumentaram o escrutínio sobre IPOs de tecnologia.

> **Por que isso importa:** A China tenta construir uma alternativa própria aos chips da NVIDIA, que enfrentam restrições dos EUA. Um IPO da T-Head daria à Alibaba dinheiro e uma vitrine para essa aposta.

Fonte original: CNBC https://www.cnbc.com/2026/11/11/alibaba-pr.html

Regulação

1 MATÉRIA NESTA EDIÇÃO

IMPORTÂNCIA ALTA 10/11/2026 · TechCrunch

Nova York abre o cadastro obrigatório da lei RAISE: OpenAI, Anthropic e outros laboratórios terão de avisar incidentes graves em 72 horas

O estado de Nova York abriu nesta terça (10/11) o portal onde os grandes desenvolvedores de IA de fronteira, empresas com mais de US$ 500 milhões de receita anual como Anthropic e OpenAI, terão de se registrar até 1º de janeiro de 2027. É quando entra em vigor a lei RAISE, que obriga essas empresas a publicar suas regras de segurança, comunicar incidentes críticos em até 72 horas e divulgar relatórios de transparência dos modelos. Um novo órgão estadual, o DIGIT, vai fiscalizar, com multas de até US$ 1 milhão na primeira violação e US$ 3 milhões nas seguintes. A governadora Kathy Hochul também lançou um centro de pesquisa em responsabilidade de IA na universidade estadual em Binghamton, que vai avaliar os sistemas de forma independente. A lei da Califórnia (SB 53) dá prazo de 15 dias para avisar incidentes.

> **Por que isso importa:** Sem uma lei federal nos EUA, são os estados que estão criando as regras para os maiores laboratórios. O prazo de 72 horas é bem mais apertado que o da Califórnia e vale para as empresas que dominam o mercado.

Fonte original: TechCrunch https://techcrunch.com/2026/11/10/openai-anthropic-face-new-oversight-as-new-yorks-ai-safety-law-takes-effect/
Infobae: Nova York exigirá aviso em 72 horas https://www.infobae.com/tecno/2026/11/11/nueva-york-exigira-a-openai-y-anthropic-informar-en-72-horas-cualquier-incidente-grave-de-seguridad-de-su-ia/

Open Source

2 MATÉRIAS NESTA EDIÇÃO

IMPORTÂNCIA MÉDIA 10/11/2026 · MarkTechPost

MBZUAI libera o PAN 2, modelo de mundo aberto de 29 bilhões de parâmetros que simula em vídeo o resultado de uma ação antes de o robô agir

O Instituto de Modelos de Fundação (IFM) da universidade MBZUAI, de Abu Dhabi, lançou nesta terça (10/11) o PAN 2, um modelo de mundo: recebe uma cena e uma ação descrita em texto e gera um vídeo de até 20 segundos com o que aconteceria. Pesos, código de treino e de inferência saíram sob a licença Apache 2.0. O sistema junta um LLM qualquer, que descreve o próximo estado do mundo, a um gerador de vídeo de 29 bilhões de parâmetros treinado do zero. Nos testes do próprio IFM, completou 34% de 100 planos de várias etapas, o melhor entre seis modelos, mas perde para o MiniMax-H3 em ações isoladas e na preferência de avaliadores humanos. Sem o LLM de planejamento, a taxa de planos completos cai para zero. Rodar exige GPUs de data center.

> **Por que isso importa:** Modelos de mundo são a aposta para treinar robôs e agentes em simulação antes de agir no mundo real. Ter um deles totalmente aberto, com código de treino, permite que qualquer laboratório estude e melhore a técnica.

Fonte original: MarkTechPost https://www.marktechpost.com/2026/11/10/ifm-pan-2-open-world-model/

IMPORTÂNCIA MÉDIA 10/11/2026 · Hugging Face

Zyphra abre o ZONOS3-Mini, voz sintética com clonagem em 43 idiomas, incluindo português, que roda em tempo real no celular e no navegador

A Zyphra publicou no Hugging Face o ZONOS3-Mini, modelo de texto para fala com pesos sob Apache 2.0. Ele tem 800 milhões de parâmetros no total, mas só cerca de 100 milhões ativos por vez, e segundo a empresa roda em tempo real em celulares e notebooks, inclusive direto no navegador via WebGPU. Clona uma voz a partir de uma amostra de áudio, permite desenhar um personagem ajustando idade, sotaque e timbre, e controlar emoção e ritmo palavra por palavra. Suporta 43 idiomas, português entre eles. Há servidores próprios para GPUs NVIDIA, Apple Silicon (MLX) e CPU, além de suporte no SGLang. A Zyphra diz medir 80 milissegundos até o primeiro áudio com 8 pedidos simultâneos numa H100.

> **Por que isso importa:** Voz de qualidade com clonagem, aberta e leve o bastante para rodar no aparelho, era um espaço dominado por serviços pagos na nuvem. Isso facilita narração, games e assistentes locais, e também aumenta a preocupação com golpes de voz clonada.

Fonte original: Hugging Face https://huggingface.co/Zyphra/ZONOS3-Mini
Discussão no r/LocalLLaMA https://www.reddit.com/r/LocalLLaMA/comments/1xtb9gw/zyphrazonos3mini_hugging_face/

Agentes e Código

3 MATÉRIAS NESTA EDIÇÃO

IMPORTÂNCIA ALTA 10/11/2026 · Ars Technica

Grok Build deixou abertos mais de 35 mil apps com dados sensíveis, inclusive documentos internos da própria SpaceXSI

A empresa de segurança Claroc achou 43.601 endereços terminados em grok.me, o domínio dos apps feitos com o Grok Build, agente de "vibe coding" da SpaceXSI (a fusão de SpaceX e xAI); mais de 35 mil, cerca de 95%, abriam para qualquer um, com todos os dados guardados. O problema é de desenho: a parte aleatória do link fica antes do domínio, e por isso aparece em bases públicas de DNS passivo que qualquer pessoa pode consultar. Entre os dados expostos havia exames médicos de 3.784 trabalhadores de uma construtora peruana, um placar de técnicos da Tesla, a lista de 108 pessoas de um contrato com a Marinha dos EUA e 24 apps da própria SpaceXSI com um plano de integração marcado como confidencial e um mapa de 594 contas de vendas. A SpaceXSI começou a tirar os links do ar depois das perguntas da Ars Technica. A Cursor, que pertence à SpaceXSI, disse que os apps são privados por padrão, o que os testes da Claroc contradizem.

> **Por que isso importa:** Quem cria um app em minutos com IA muitas vezes não pensa em quem mais pode abri-lo. O caso mostra que o "compartilhar com quem tem o link" só protege se o link não vazar, e a Claroc diz que Lovable, Bolt e Replit têm o mesmo tipo de exposição.

Fonte original: Ars Technica https://arstechnica.com/security/2026/11/grok-build-coding-agent-leaks-oodles-of-confidential-info-some-belonging-to-spacexsi/
Techmeme: repercussão https://www.techmeme.com/261111/p3#a261111p3

IMPORTÂNCIA MÉDIA 11/11/2026 · IT之家 (IT Home)

Alibaba lança o Qwen Code 1.0 Preview e abre, junto, o núcleo de agentes do Qoder para quem quiser embutir em produtos próprios

A Alibaba anunciou nesta quarta (11/11) o Qwen Code 1.0, que incorpora ao projeto de código aberto Qwen Code o "harness" (a camada que liga modelo, ferramentas e execução) desenvolvido no Qoder, sua ferramenta de programação com agentes. A ideia é oferecer uma base de agente que desenvolvedores possam embutir nos próprios produtos, com um modelo de eventos unificado que separa o raciocínio da execução de ferramentas e caminha para rodar em vários nós e sessões. A versão sai num ramo separado de prévia, e o ramo estável continua. Segundo as notas no GitHub, a prévia aceita contas Qwen e Qoder, vários provedores de modelo e endpoints compatíveis, traz pacotes nativos para macOS, Linux e Windows sem precisar de Node.js e permite retomar sessões e rodar sem interação.

> **Por que isso importa:** Coloca na mesa um concorrente aberto, sob Apache 2.0, para os motores de agentes de código como o do Claude Code. Para quem constrói agentes, é uma base pronta e usada em produção que não depende de um fornecedor fechado.

Fonte original: IT之家 (IT Home) https://www.ithome.com/1/025/393.htm
Qwen Code no GitHub https://github.com/QwenLM/qwen-code

IMPORTÂNCIA BAIXA 10/11/2026 · JetBrains

ReSharper 2026.3 coloca Claude Agent e Codex dentro do Visual Studio, ao lado do Junie, no dia do lançamento do .NET 11

A JetBrains lançou nesta terça (10/11) o ReSharper 2026.3, que deixa escolher Claude Agent, Codex ou Junie na janela de IA do Visual Studio para escrever e editar código, rodar comandos e explicar partes do projeto. Os agentes são ligados pelo Agent Client Protocol (ACP), um protocolo aberto, e podem usar as refatorações do próprio ReSharper, como renomear símbolos. O uso de Claude Agent e Codex consome a cota da assinatura JetBrains AI; não dá para usar assinatura própria do Claude ou do ChatGPT nem chave de API. Há um nível gratuito do Junie, o Junie Lite. A versão também suporta o .NET 11 e o C# 15, lançados no mesmo dia na .NET Conf 2026.

> **Por que isso importa:** Leva os agentes de código mais usados para o editor de muitas equipes .NET, inclusive em bancos e grandes empresas, sem trocar de ferramenta.

Fonte original: JetBrains https://blog.jetbrains.com/dotnet/2026/11/10/resharper-2026-3-release/

Modelos Locais

3 MATÉRIAS NESTA EDIÇÃO

IMPORTÂNCIA MÉDIA 10/11/2026 · GitHub (ggml-org/llama.cpp)

llama.cpp passa a aceitar cache KV em FP8, o que abre caminho para rodar localmente modelos publicados nesse formato

O llama.cpp, motor mais usado para rodar modelos no próprio computador, incorporou nesta terça (10/11) o suporte a cache KV (a memória de atenção que cresce com o tamanho da conversa) no formato OCP FP8, com escalas por tensor opcionais. A mudança, de 710 linhas, foi aprovada por Georgi Gerganov e entrou na versão b12274. O autor cita o Mistral-Medium-3.5-128B em NVFP4, da NVIDIA, como exemplo de modelo que usa esse tipo de cache, e informou ter feito o trabalho com ajuda do Codex. O suporte para modelos com MLA e DFlash ficou para outra proposta, ainda em rascunho.

> **Por que isso importa:** Quem roda modelos grandes em casa ou num DGX esbarra na memória do cache em conversas longas. Com o FP8, checkpoints pensados para vLLM passam a funcionar no llama.cpp sem trocar de motor.

Fonte original: GitHub (ggml-org/llama.cpp) https://github.com/ggml-org/llama.cpp/pull/28392
Versão b12274 https://github.com/ggml-org/llama.cpp/releases/tag/b12274
Discussão no r/LocalLLaMA https://www.reddit.com/r/LocalLLaMA/comments/1xt9o3t/llamacpp_add_fp8_kvcache/

IMPORTÂNCIA MÉDIA 10/11/2026 · MacRumors

Claris, da Apple, testa o Harbor, app que transforma Macs em servidor de IA local e junta vários Macs pelo Thunderbolt 5

A Claris, subsidiária da Apple conhecida pelo FileMaker, está testando o Harbor, app que transforma um Mac com chip Apple num servidor privado de inferência para modelos de pesos abertos, sem mandar nada para a nuvem. Segundo a descrição publicada, ele detecta sozinho outros Macs ligados por Thunderbolt 5 e soma a memória unificada deles para rodar modelos maiores do que caberiam em um só, sem linha de comando nem configuração manual; também deixa buscar, baixar e carregar modelos dentro do app. O Harbor está em acesso antecipado; a página na App Store chegou a aparecer e foi retirada. O cadastro pede o nome da empresa, o que sugere foco corporativo.

> **Por que isso importa:** Seria a primeira ferramenta da própria Apple para IA local em cluster, um espaço hoje ocupado por projetos como Exo e LM Studio. Para empresas com dados sensíveis, é um caminho para rodar modelos fortes sem tirar dados de casa.

Fonte original: MacRumors https://www.macrumors.com/2026/11/10/apple-claris-harbor-app/

IMPORTÂNCIA BAIXA 11/11/2026 · TechNode

Huawei explica como o Kirin 9050 roda no celular, sem internet, um modelo de 30 bilhões de parâmetros

A Huawei detalhou nesta quarta (11/11) como os chips Kirin 9050 rodam localmente, sem conexão, um modelo de IA de 30 bilhões de parâmetros. A nova unidade de processamento neural Da Vinci usa quantização de poucos bits e coordena memória e computação para caber no telefone. O modelo é do tipo mistura de especialistas e ativa cerca de 2 bilhões de parâmetros por vez, número que a empresa já tinha revelado em outubro. Ela também mostrou recursos que o processador sustenta, como leitura de tela por IA e virar a página com o olhar.

> **Por que isso importa:** Mostra quanto a IA no próprio aparelho avançou, e que a China segue investindo em chips próprios para celular apesar das restrições dos EUA.

Fonte original: TechNode https://technode.com/2026/11/11/huawei-details-how-kirin-9050-runs-a-30-billion-parameter-model-on-device/

Finanças

1 MATÉRIA NESTA EDIÇÃO

IMPORTÂNCIA MÉDIA 11/11/2026 · Finextra

McKinsey: IA pode cortar 20% ou mais dos custos dos bancos, mas 74% de quem usa IA para finanças recorre a ChatGPT, Claude e Gemini, não ao banco

O relatório anual da McKinsey sobre bancos, "Global Banking Annual Review 2026: The AI disruption dividend", diz que a IA pode reduzir em 20% ou mais a base de custos do setor, mas que a maioria das instituições só a usou até agora para aumentar a produtividade individual, com efeito desprezível no resultado. Numa pesquisa com 46 mil pessoas em 13 países, 23% disseram usar IA generativa com frequência para tarefas bancárias, como comparar produtos e pedir conselhos; desses, 74% usam plataformas de terceiros como ChatGPT, Claude e Gemini, contra 37% que usam o chatbot do próprio banco. Ainda assim, 62% prefeririam receber serviços com IA do seu banco principal, e 72% deixariam um agente fazer ao menos uma tarefa financeira por eles, a maioria com pedido de permissão antes; cerca de 65% limitariam aprovações automáticas a US$ 500.

> **Por que isso importa:** Os números mostram que os assistentes de IA estão virando uma camada entre o banco e o cliente. Para bancos como os brasileiros, a corrida é oferecer agentes próprios confiáveis antes que o cliente se acostume a resolver tudo fora do app.

Fonte original: Finextra https://www.finextra.com/newsarticle/48739/ai-could-cut-banking-industrys-cost-base-by-20---mckinseys

Games

2 MATÉRIAS NESTA EDIÇÃO

IMPORTÂNCIA MÉDIA 10/11/2026 · CNET

Metade dos jogos lançados na Steam em outubro declarou uso de IA generativa, mas eles somam menos avaliações e notas piores

Um levantamento do usuário Ploz, no Reddit, com os dados públicos da Steam mostra que 49,9% dos jogos novos lançados em outubro declararam uso de IA em conteúdo visível ao jogador, como arte, som, diálogos e tradução. A declaração é obrigatória desde janeiro de 2024, quando o índice era de 1,4%; subiu para 13,4% no fim de 2024 e 24% no fim de 2025. O uso mais comum é em arte (63% das declarações), seguido de texto e tradução (30%), música (21%), código (16%) e vozes (7%). Os jogos com IA recebem, em média, avaliações menos positivas, e só 17% dos lançamentos com mais de mil avaliações têm a declaração. A Valve não comentou; a CNET procurou duas dúzias de grandes estúdios, e todos se recusaram a comentar.

> **Por que isso importa:** Mostra o tamanho da enxurrada de jogos feitos com IA nas lojas, e que o público ainda prefere os feitos sem ela. Para estúdios pequenos, fica mais difícil aparecer no meio da multidão.

Fonte original: CNET https://www.cnet.com/tech/gaming/50-percent-new-steam-games-ai/

IMPORTÂNCIA BAIXA 10/11/2026 · Kotaku

Competição australiana de criação de jogos para estudantes acaba depois de 12 anos porque ficou impossível saber o que foi feito com IA

O Australian STEM Videogame Challenge, aberto a estudantes do ensino fundamental ao médio, terá em 2026 sua última edição. Segundo a organizadora ACER, citada pelo PC Gamer e pelo Kotaku, "o rápido e complexo avanço da IA generativa no desenvolvimento de jogos" tornou mais difícil julgar o que os alunos fizeram sozinhos e afastou o programa do seu propósito original. No Reddit, a mãe ou o pai de um participante contou que o filho, que competia havia anos com os amigos, ficou "visivelmente arrasado" com a notícia.

> **Por que isso importa:** É um efeito colateral pouco discutido da IA: espaços de aprendizado para crianças estão sumindo porque não há como separar o trabalho delas do da máquina.

Fonte original: Kotaku https://kotaku.com/student-game-development-competition-shut-down-because-kids-kept-submitting-ai-slop-games-2000755605

Brasil

3 MATÉRIAS NESTA EDIÇÃO

IMPORTÂNCIA MÉDIA 11/11/2026 · Forbes Brasil

Brasil cai duas posições e fica em 54º entre 70 países no ranking mundial de digital e IA do IMD com a Fundação Dom Cabral

O Brasil ficou em 54º lugar entre 70 economias no World Digital and AI Ranking 2026, feito pelo IMD, escola de negócios suíça, com a Fundação Dom Cabral, duas posições abaixo da edição anterior. O país vai bem em infraestrutura de pesquisa (26º) e liderança pública (28º), mas aparece em 65º em adoção de IA e 66º em capital e investimento na tecnologia. Os Estados Unidos lideram, seguidos de Singapura e Holanda; a China subiu sete posições, para 7º. No ranking de empresas, a NVIDIA lidera pelo segundo ano. Para Hugo Tadeu, diretor de inovação da FDC, o resultado "não é terra arrasada", mas o país precisa avançar em gestão, medir o retorno da IA e formar mais engenheiros. O ranking avalia 88 critérios, 51 estatísticos e 37 de pesquisa com executivos.

> **Por que isso importa:** O Brasil está entre os maiores usuários de ferramentas como o ChatGPT, mas o ranking mostra que isso não está virando investimento e produtividade nas empresas.

Fonte original: Forbes Brasil https://forbes.com.br/forbes-money/2026/11/brasil-ranking-global-ia-imd-fdc/
Época Negócios: Brasil em 54º lugar https://epocanegocios.globo.com/inteligencia-artificial/noticia/2026/11/brasil-fica-em-54o-lugar-em-ranking-global-de-inteligencia-artificial-com-70-paises-nvidia-lidera-entre-empresas.ghtml

IMPORTÂNCIA MÉDIA 11/11/2026 · Época Negócios

OpenAI diz que o hábito brasileiro de mandar áudios ajuda a melhorar sua voz e que o Brasil é o 2º país em desenvolvedores na API

Em entrevista à Época Negócios, Matias Castello, líder de experiência para desenvolvedores da OpenAI, contou que a empresa usa o retorno de mercados onde mandar áudio é comum, como o Brasil, para aperfeiçoar suas ferramentas de voz, incluindo novas vozes e sotaques. Segundo dados apresentados pela companhia, o país soma cerca de 215 milhões de mensagens por dia no ChatGPT, o português é o terceiro idioma mais usado e o Brasil está entre os cinco maiores mercados em geração de imagens e assinaturas. O Brasil também é o segundo país em número de desenvolvedores que usam a API, atrás só dos EUA, e o maior mercado do Codex na América Latina, com usuários multiplicados por dez a onze vezes desde o início de 2026. Castello disse ainda que a OpenAI tirou engenheiros de produtos novos para reforçar segurança e alinhamento.

> **Por que isso importa:** Os números, da própria OpenAI, mostram o peso do Brasil para a empresa justamente quando ela troca todos os usuários para um novo modelo de voz.

Fonte original: Época Negócios https://epocanegocios.globo.com/inteligencia-artificial/noticia/2026/11/como-o-habito-brasileiro-de-mandar-audios-ajudou-a-openai-a-melhorar-sua-tecnologia-de-voz.ghtml

IMPORTÂNCIA BAIXA 10/11/2026 · Olhar Digital

99Food troca fotos reais de restaurantes por imagens feitas com IA sem autorização e diz que está revisando a moderação dos cardápios

Restaurantes relatam, nas redes e no Reclame Aqui, pelo menos desde abril, que a 99Food substitui as fotos reais dos pratos por imagens criadas ou modificadas com IA sem pedir autorização, e que volta a trocá-las depois que os donos recuperam as originais. A plataforma diz que está revisando seus critérios de moderação de cardápios. O contrato permite editar ou substituir fotos com IA, mas exige avisar o estabelecimento e identificar as imagens de forma visível para o consumidor, e deixa o restaurante removê-las quando quiser. Uma chocolateria de Santo André (SP) contou à Folha que reclama desde julho e que, quando o cliente recebe algo diferente da foto, é a loja que paga o estorno.

> **Por que isso importa:** É um exemplo bem brasileiro de IA usada para "melhorar" a vitrine sem combinar com quem vende, e de como o risco acaba ficando com o pequeno comerciante.

Fonte original: Olhar Digital https://olhardigital.com.br/2026/11/10/internet-e-redes-sociais/99food-altera-fotos-de-restaurantes-com-ia-sem-autorizacao-e-enfrenta-reclamacoes/

Fofocas das redes

4 MATÉRIAS NESTA EDIÇÃO

X IMPORTÂNCIA MÉDIA 10/11/2026 · @willdepue

Ex-pesquisador da OpenAI responde a uma colega que "acabou" para o trabalho humano não artesanal e compara a situação a Magnus Carlsen contra o Stockfish

Will Depue, que trabalhou na OpenAI, respondeu a uma pesquisadora que tentava passar uma visão otimista a estudantes preocupados com o futuro. Ele escreveu que está "totalmente acabado" para a raça humana e para todo trabalho econômico que não tira seu valor de ser feito por uma pessoa, que em breve todos vão assistir "da arquibancada" e que daqui em diante é "Magnus Carlsen contra o Stockfish 13". A pesquisadora tinha dito que acredita que "o bolo está crescendo para todo mundo".

> **Por que isso importa:** O tom fatalista de alguém que esteve dentro de um laboratório de ponta alimentou um dos debates mais ativos do dia no X sobre emprego e propósito na era da IA.

Engajamento: 172,6 mil visualizações · 1,2 mil curtidas · 42 reposts · 64 citações · 77 respostas · 460 salvamentos (medido em 11/11/2026 07:40)
Post original (X): @willdepue https://x.com/willdepue/status/2120282720853197236
Post da pesquisadora a quem ele respondeu https://x.com/niloofar_mire/status/2120281216259870817

X IMPORTÂNCIA BAIXA 10/11/2026 · @nerdalert

"Uma alucinação de LLM fez 72 ovos chegarem à minha casa": post sobre compra feita por IA vira piada e debate

Matt Smith (@nerdalert) postou que uma alucinação de um modelo de linguagem resultou na entrega de 72 ovos na casa dele, com uma foto. Nas respostas, gente sugeriu receitas e fez piada com cenários piores, enquanto outros criticaram quem deixa a IA fazer as compras: "o tempo que você leva para dizer à IA o que quer, você já teria feito o pedido", escreveu uma usuária.

> **Por que isso importa:** É um exemplo leve, mas concreto, do que acontece quando agentes passam a fazer compras em nome das pessoas sem uma conferência humana.

Engajamento: 127,9 mil visualizações · 1,6 mil curtidas · 70 reposts · 58 citações · 245 respostas · 172 salvamentos (medido em 11/11/2026 07:40)
Post original (X): @nerdalert https://x.com/nerdalert/status/2120251385199419805
Crítica de @MxVivianWulf https://x.com/MxVivianWulf/status/2120412184714842477

X IMPORTÂNCIA BAIXA 11/11/2026 · @waitbutwhy

Tim Urban, do Wait But Why, resume o vaivém de quem acompanha IA: “estamos ferrados” lendo as notícias, “nada mudou” saindo na rua

Tim Urban, autor do blog Wait But Why, publicou um diálogo consigo mesmo: ao ler sobre o que está acontecendo com a IA, conclui que “estamos realmente ferrados”; ao sair na rua e ver todo mundo agindo normalmente, acha que estava numa bolha paranoica; ao voltar a ler, decide que os “delirantes” são as pessoas lá fora, e o ciclo recomeça. O post ganhou réplicas com comparações como banhistas ignorando um tsunami e críticos lembrando que outras ondas de tecnologia trouxeram crescimento.

> **Por que isso importa:** Traduz em humor a distância entre quem acompanha a IA de perto e o resto da população, tema recorrente nas redes nesta semana de previsões cada vez mais curtas.

Engajamento: 135,1 mil visualizações · 4,8 mil curtidas · 336 reposts · 60 citações · 174 respostas · 674 salvamentos (medido em 11/11/2026 07:40)
Post original (X): @waitbutwhy https://x.com/waitbutwhy/status/2120392816681492842

REDDIT IMPORTÂNCIA MÉDIA 11/11/2026 · r/LocalLLaMA

Boato no r/LocalLLaMA: NVIDIA teria cancelado a RTX 5080 de 24 GB, a placa que quem roda IA em casa esperava

Um post no r/LocalLLaMA levou ao topo do dia uma reportagem do VideoCardz, marcada pelo próprio site como boato, segundo a qual a NVIDIA avisou parceiros na China que a GeForce RTX 5080 com 24 GB de memória não será mais lançada em janeiro de 2027 e que a versão de 16 GB volta a ser produzida. O motivo seria o preço dos chips GDDR7 de 3 GB, cerca de três vezes mais caros que os de 2 GB. A NVIDIA não confirmou nem a placa nem o cancelamento.

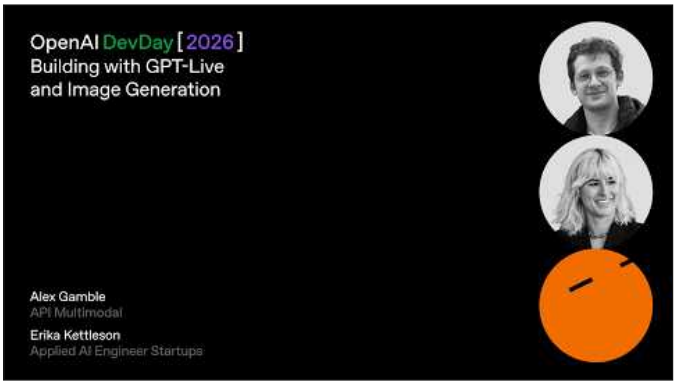
> **Por que isso importa:** Memória de vídeo é o que limita o tamanho dos modelos que dá para rodar em casa, e 24 GB numa placa de linha seria um alívio para essa comunidade. A falta de memória barata segue travando o hardware para IA local.

Post original (Reddit): r/LocalLLaMA https://www.reddit.com/r/LocalLLaMA/comments/1xtdwps/nvidia_reportedly_cancels_geforce_rtx_5080_24gb/
VideoCardz: reportagem original (boato) https://videocardz.com/newz/nvidia-reportedly-cancels-geforce-rtx-5080-24gb-launch-brings-back-16gb-model-over-gddr7-costs

Vídeos

4 VÍDEOS

Os resumos se baseiam no título verificado e nos anúncios relacionados lidos para esta edição. Clique na miniatura ou no título para abrir no YouTube.



Building with GPT-Live and Image Generation | DevDay 2026

OpenAI · 10/11/2026

Sessão do DevDay 2026 em que a OpenAI mostra o GPT-Live conversando, jogando e criando imagens enquanto coordena tarefas em segundo plano, e ensina a pôr voz em agentes lidando com atraso, interrupções e mudanças de pedido.

https://www.youtube.com/watch?v=jDxKzlGLCas

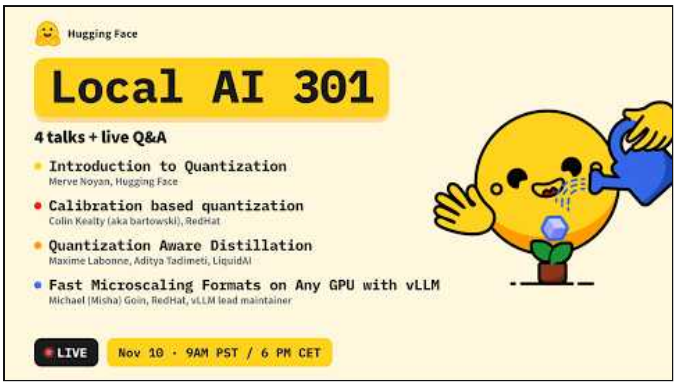


Fable 5.2: The Self-Improvement Loop Keeps Loopin' ... (plus Gemini Carbon)

AI Explained · 10/11/2026

O canal AI Explained diz ter lido o relatório de 264 páginas do Fable 5.2 e o relatório do METR e testado o Fable 5.2 e o Gemini Carbon centenas de vezes, e discute o quanto a IA já acelera a própria pesquisa de IA.

https://www.youtube.com/watch?v=583ThJIqQ7k



Local AI 301: Quantization

Hugging Face · 11/11/2026

Transmissão do Hugging Face sobre quantização, com Merve Noyan, Colin Kealty (bartowski) e Michael Goin, da Red Hat, e Maxime Labonne, da Liquid AI, mostrando como começar com llama.cpp e vLLM.

https://www.youtube.com/watch?v=il0dgxjpoC0



Ask the Experts: Dory Looped Transformer Model | Nemotron Labs

NVIDIA Developer · 11/11/2026

Pesquisadores da NVIDIA explicam o Dory 1.0, modelo aberto de 4 bilhões de parâmetros que repete um bloco de camadas para trocar mais computação por mais acerto: segundo eles, sobe de 50,3 para 61,8 no SWE-bench Verified com quatro voltas.

https://www.youtube.com/watch?v=_MEqhFwnp8o

SOBRE ESTA EDIÇÃO
Nº 23 — Edição da manhã · 11/11/2026

JANELA
Cobre o que saiu desde a edição nº 22 (fechada por volta das 18h de Brasília de 10/11) até por volta das 7h45 de Brasília de 11/11, incluindo a checagem do vigia horário das 7h21. A manchete é o processo de acionista contra a NVIDIA e a contratação do escritório Wachtell Lipton para investigar o caso de Jensen Huang. A edição traz ainda o Gemini Robotics 2.5, a técnica Lazarus da Microsoft Research, o fim dos modos de voz antigos do ChatGPT, o investimento da TSMC na OpenAI e na Anthropic, a receita trimestral de US$ 10 bilhões da OpenAI, a rodada da Mercor, a compra da Odessia, o IPO da T-Head, a lei RAISE de Nova York, o PAN 2, o ZONOS3-Mini, o vazamento do Grok Build, o Qwen Code 1.0, o ReSharper 2026.3, o FP8 no llama.cpp, o Harbor da Apple, o Kirin 9050, o relatório da McKinsey sobre bancos, a Steam com metade dos lançamentos feitos com IA, o fim do STEM Videogame Challenge, o ranking de IA do IMD com a FDC, a OpenAI e o Brasil e a polêmica das fotos da 99Food. Fechamento: 11/11/2026, por volta das 7h45 de Brasília.

FONTES VARRIDAS
Blogs e changelogs oficiais (OpenAI, Anthropic, Google DeepMind, Meta, xAI, Microsoft, NVIDIA, Apple, Mistral, DeepSeek, Qwen, Moonshot), arXiv e HF Papers, trending do Hugging Face e do GitHub, X, imprensa dos EUA, Europa, China, Japão, Índia e Brasil, e YouTube.

CRITÉRIO
Só entram notícias lidas numa fonte real, de preferência primária, com todos os links checados. Páginas com bloqueio (403) foram mantidas; links quebrados (404) foram descartados.

Fontes e links

▶ **VÍDEOS SUGERIDOS**

1. **Building with GPT-Live and Image Generation | DevDay 2026** — OpenAI, 10/11/2026
   https://www.youtube.com/watch?v=jDxKzlGLCas
2. **Fable 5.2: The Self-Improvement Loop Keeps Loopin' ... (plus Gemini Carbon)** — AI Explained, 10/11/2026
   https://www.youtube.com/watch?v=583ThJIqQ7k
3. **Local AI 301: Quantization** — Hugging Face, 11/11/2026
   https://www.youtube.com/watch?v=il0dgxjpoC0
4. **Ask the Experts: Dory Looped Transformer Model | Nemotron Labs** — NVIDIA Developer, 11/11/2026
   https://www.youtube.com/watch?v=_MEqhFwnp8o

Manchete

NVIDIA shareholders are suing the company and CEO Jensen Huang — Engadget
https://www.engadget.com/2312947/nvidia-shareholders-are-suing-the-company-and-ceo-jensen-huang/

Nvidia hires Wachtell Lipton for Jensen Huang investigation — Politico Europe
https://www.politico.eu/article/nvidia-hires-wachtell-lipton-for-jensen-huang-investigation/

Nvidia shares rebound 4% as Huang faces scrutiny ahead of earnings — CNBC
https://www.cnbc.com/2026/11/11/nvidia-stock-jensen-huang-affair-china.html

Modelos

Gemini Robotics 2.5: robô humanoide com o novo modelo do Google DeepMind separa peças em linhas da Mercedes-Benz com 96% de acerto
Google DeepMind: https://deepmind.google/blog/gemini-robotics-2-5-accelerates-real-world-physical-intelligence/
Ars Technica: Gemini Robotics 2.5 e o teste ASIMOV-jailbreak: https://arstechnica.com/google/2026/11/deepmind-reveals-gemini-robotics-2-5-and-open-safety-benchmark-asimov-jailbreak/

Pesquisa

Microsoft Research mostra que editar menos de 10 neurônios "adormecidos" basta para plantar um backdoor num agente de IA, sem treinar o modelo
Microsoft Research: https://www.microsoft.com/en-us/research/blog/lazare-veni-foras-reprogramming-dormant-neurons-to-backdoor-ai-agents/
Discussão no r/LocalLLaMA: https://www.reddit.com/r/LocalLLaMA/comments/1xt82mz/microsoft_research_backdoored_an_ai_agent_by/

Produtos

OpenAI vai desligar os modos de voz Avançado e Padrão do ChatGPT em 11/12 e levar todos ao GPT-Live, e parte dos usuários reclama
Neowin: https://www.neowin.net/news/openai-is-retiring-chatgpts-advanced-and-standard-voice-modes-sparking-user-backlash/

Negócios e Investimento

TSMC aprova investir até US$ 500 milhões na OpenAI e outros US$ 500 milhões na Anthropic, sua primeira aposta direta em laboratórios de IA
BusinessKorea: https://www.businesskorea.co.kr/news/articleView.html?idxno=281064
SCMP: o que a TSMC ganha com as participações: https://www.scmp.com/tech/big-tech/article/3374037/what-taiwanese-chip-giant-tsmc-gains-taking-stakes-openai-and-anthropics-ai-race

OpenAI fatura US$ 10 bilhões no terceiro trimestre, cerca de 50% a mais, mas o prejuízo operacional sobe para US$ 10,9 bilhões, diz o WSJ
Reuters (via The Hindu): https://www.thehindu.com/sci-tech/technology/openai-posts-steep-losses-even-as-quarterly-revenue-rises-to-10-billion-report/article71688191.ece

Mercor levanta US$ 750 milhões a US$ 20 bilhões de avaliação para levar a avaliação de agentes de IA às empresas
SiliconANGLE: https://siliconangle.com/2026/11/10/mercor-raises-750m-at-20b-valuation-to-take-ai-agent-evaluation-to-enterprises/

OpenAI compra a Odessia, agente de viagens com IA criado pelo fundador da Sonder, e leva o fundador para cuidar do "futuro das viagens"
TechCrunch: https://techcrunch.com/2026/11/10/openai-acquires-odessia-travel-startup-created-by-sonder-founder/
PYMNTS: ações de agências de viagem caem: https://www.pymnts.com/news/artificial-intelligence/2026/online-travel-agencies-shares-drop-following-openai-acquisition-of-travel-startup-odessia/

Alibaba anuncia que vai abrir o capital da T-Head, sua unidade de chips de IA que disputa espaço com a NVIDIA na China, em Hong Kong
CNBC: https://www.cnbc.com/2026/11/11/alibaba-pr.html

Regulação

Nova York abre o cadastro obrigatório da lei RAISE: OpenAI, Anthropic e outros laboratórios terão de avisar incidentes graves em 72 horas
TechCrunch: https://techcrunch.com/2026/11/10/openai-anthropic-face-new-oversight-as-new-yorks-ai-safety-law-takes-effect/
Infobae: Nova York exigirá aviso em 72 horas: https://www.infobae.com/tecno/2026/11/11/nueva-york-exigira-a-openai-y-anthropic-informar-en-72-horas-cualquier-incidente-grave-de-seguridad-de-su-ia/

Open Source

MBZUAI libera o PAN 2, modelo de mundo aberto de 29 bilhões de parâmetros que simula em vídeo o resultado de uma ação antes de o robô agir
MarkTechPost: https://www.marktechpost.com/2026/11/10/ifm-pan-2-open-world-model/

Zyphra abre o ZONOS3-Mini, voz sintética com clonagem em 43 idiomas, incluindo português, que roda em tempo real no celular e no navegador
Hugging Face: https://huggingface.co/Zyphra/ZONOS3-Mini
Discussão no r/LocalLLaMA: https://www.reddit.com/r/LocalLLaMA/comments/1xtb9gw/zyphrazonos3mini_hugging_face/

Agentes e Código

Grok Build deixou abertos mais de 35 mil apps com dados sensíveis, inclusive documentos internos da própria SpaceXSI
Ars Technica: https://arstechnica.com/security/2026/11/grok-build-coding-agent-leaks-oodles-of-confidential-info-some-belonging-to-spacexsi/
Techmeme: repercussão: https://www.techmeme.com/261111/p3#a261111p3

Alibaba lança o Qwen Code 1.0 Preview e abre, junto, o núcleo de agentes do Qoder para quem quiser embutir em produtos próprios
IT之家 (IT Home): https://www.ithome.com/1/025/393.htm
Qwen Code no GitHub: https://github.com/QwenLM/qwen-code

ReSharper 2026.3 coloca Claude Agent e Codex dentro do Visual Studio, ao lado do Junie, no dia do lançamento do .NET 11
JetBrains: https://blog.jetbrains.com/dotnet/2026/11/10/resharper-2026-3-release/

Modelos Locais

llama.cpp passa a aceitar cache KV em FP8, o que abre caminho para rodar localmente modelos publicados nesse formato
GitHub (ggml-org/llama.cpp): https://github.com/ggml-org/llama.cpp/pull/28392
Versão b12274: https://github.com/ggml-org/llama.cpp/releases/tag/b12274
Discussão no r/LocalLLaMA: https://www.reddit.com/r/LocalLLaMA/comments/1xt9o3t/llamacpp_add_fp8_kvcache/

Claris, da Apple, testa o Harbor, app que transforma Macs em servidor de IA local e junta vários Macs pelo Thunderbolt 5
MacRumors: https://www.macrumors.com/2026/11/10/apple-claris-harbor-app/

Huawei explica como o Kirin 9050 roda no celular, sem internet, um modelo de 30 bilhões de parâmetros
TechNode: https://technode.com/2026/11/11/huawei-details-how-kirin-9050-runs-a-30-billion-parameter-model-on-device/

Finanças

McKinsey: IA pode cortar 20% ou mais dos custos dos bancos, mas 74% de quem usa IA para finanças recorre a ChatGPT, Claude e Gemini, não ao banco
Finextra: https://www.finextra.com/newsarticle/48739/ai-could-cut-banking-industrys-cost-base-by-20---mckinseys

Games

Metade dos jogos lançados na Steam em outubro declarou uso de IA generativa, mas eles somam menos avaliações e notas piores
CNET: https://www.cnet.com/tech/gaming/50-percent-new-steam-games-ai/

Competição australiana de criação de jogos para estudantes acaba depois de 12 anos porque ficou impossível saber o que foi feito com IA
Kotaku: https://kotaku.com/student-game-development-competition-shut-down-because-kids-kept-submitting-ai-slop-games-2000755605

Brasil

Brasil cai duas posições e fica em 54º entre 70 países no ranking mundial de digital e IA do IMD com a Fundação Dom Cabral
Forbes Brasil: https://forbes.com.br/forbes-money/2026/11/brasil-ranking-global-ia-imd-fdc/
Época Negócios: Brasil em 54º lugar: https://epocanegocios.globo.com/inteligencia-artificial/noticia/2026/11/brasil-fica-em-54o-lugar-em-ranking-global-de-inteligencia-artificial-com-70-paises-nvidia-lidera-entre-empresas.ghtml

OpenAI diz que o hábito brasileiro de mandar áudios ajuda a melhorar sua voz e que o Brasil é o 2º país em desenvolvedores na API
Época Negócios: https://epocanegocios.globo.com/inteligencia-artificial/noticia/2026/11/como-o-habito-brasileiro-de-mandar-audios-ajudou-a-openai-a-melhorar-sua-tecnologia-de-voz.ghtml

99Food troca fotos reais de restaurantes por imagens feitas com IA sem autorização e diz que está revisando a moderação dos cardápios
Olhar Digital: https://olhardigital.com.br/2026/11/10/internet-e-redes-sociais/99food-altera-fotos-de-restaurantes-com-ia-sem-autorizacao-e-enfrenta-reclamacoes/

Fofocas das redes

Ex-pesquisador da OpenAI responde a uma colega que "acabou" para o trabalho humano não artesanal e compara a situação a Magnus Carlsen contra o Stockfish
@willdepue: https://x.com/willdepue/status/2120282720853197236
Post da pesquisadora a quem ele respondeu: https://x.com/niloofar_mire/status/2120281216259870817

"Uma alucinação de LLM fez 72 ovos chegarem à minha casa": post sobre compra feita por IA vira piada e debate
@nerdalert: https://x.com/nerdalert/status/2120251385199419805
Crítica de @MxVivianWulf: https://x.com/MxVivianWulf/status/2120412184714842477

Tim Urban, do Wait But Why, resume o vaivém de quem acompanha IA: "estamos ferrados" lendo as notícias, "nada mudou" saindo na rua
@waitbutwhy: https://x.com/waitbutwhy/status/2120392816681492842

Boato no r/LocalLLaMA: NVIDIA teria cancelado a RTX 5080 de 24 GB, a placa que quem roda IA em casa esperava
r/LocalLLaMA: https://www.reddit.com/r/LocalLLaMA/comments/1xtdwps/nvidia_reportedly_cancels_geforce_rtx_5080_24gb/
VideoCardz: reportagem original (boato): https://videocardz.com/newz/nvidia-reportedly-cancels-geforce-rtx-5080-24gb-launch-brings-back-16gb-model-over-gddr7-costs

48 links únicos desta edição, sem repetir URLs.