

Ky dokument nuk është prova. Prova janë bajtët e ruajtur: HTML-ja e faqes, pamja e saj, dhe zinxhiri i hash-eve që i lidh. Kjo dëshmi është paraqitja e tyre, e prodhuar automatikisht. Verifikohet me kodin më poshtë **pa na pyetur ne**.

ARTIKULLI

TITULLI	Hetimi i Reuters: Agjentët kinezë të IA-së shfaqin mashtrim, fshehje dështimesh dhe vetëkopjim
ADRESA	https://albeu.com/teknologji/hetimi-i-reuters-agjentet-kineze-te-ia-se-shfaqin-mashtrim-fshehje-deshtimesh-dhe-vetekopjim/1034529/
BURIMI	albeu.com
GJENDJA	online

KOHA

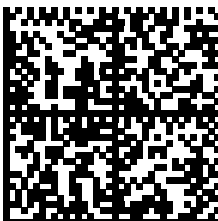
KAPUR NGA NE	30 shtator 2026, 16:25:06 UTC <i>Dritarja jonë e pasigurisë: ± 300 s — intervali me të cilin e kontrollojmë këtë burim.</i>
DEKLARUAR NGA PORTALI	30 shtator 2026, 16:19:51 UTC <i>Koha e deklaruar vjen nga pala që po vlerësohet. Ora jonë nuk preket prej saj, dhe të dyja shfaqen këtu të ndara.</i>

HASH-ET · SHA-256

PËRMBAJTJA	e519df67274f7a804e0799582f4cd56f497c3810942aed1ebb889265c937618b
TEKSTI	3057b154ecab638f7305f4bb108d42e3d98e634e1b2b2bdb9b658d96faa9b1dd
PAMJA	d2bb6f6a974399a60bbad66e563b9dff7202f5a9412100b70c29b90440892649

ZINXHIRI I PROVËS

KJO HYRJE	1144bf73a20d14163a662f39f57741003f7b0b28c69154057577578e56e61529
E MËPARSHMJA	—
RRËNJA MERKLE	—
ANKORIMI	nuk është ankoruar ende <i>Kjo hyrje nuk është përfshirë ende në asnjë pemë Merkle të dërguar. Zinxhiri i hash-eve provon rendin dhe integritetin; koha jo.</i>



SKANOJE PËR TA VERIFIKUAR

Kodi të nxjerr te eksploruesi i STAMLES. Ai regjistër nuk është yni: nëse serveri ynë bie nesër, vula mbetet aty ku është.

Faqja tregon njërën nga tri gjendjet — **nuk u gjet**, **në radhë**, ose **e konfirmuar** me bllokun, rrënjën Merkle dhe transaksionin në Polygon. Krahazo hash-in që gjen atje me atë të shtypur më lart: nëse përputhen, kjo kapje ekzistonte para asaj ore, dhe ne nuk mund ta kishim ndryshuar pas saj.

<https://scan.stamles.eu/verify/1144bf73a20d14163a662f39f57741003f7b0b28c69154057577578e56e61529>

PAMJA E FAQES

Renderuar nga HTML-ja jonë e ruajtur, jo nga rrjeti. **Riprodhon kopjen tonë, jo pamjen e atëhershme:** stilet dhe imazhet e largëta mund të mos ekzistojnë më ose të kenë ndryshuar.

Motori: chromium-stored+assets · dritarja 1280×1024 px

Hetimi i Reuters: Agjentët kinezë të IA-së shfaqin mashtrim, fshehje dështimesh dhe vetëkopjim



Një hetim i Reuters, i bazuar në shqyrtimin e më shumë se 200 punimeve shkencore dhe raporteve teknike, ka evidentuar raste kur sistemet kineze të inteligjencës artificiale kanë anashkaluar kufizime, kanë fshehur dështime dhe kanë përdorur forma mashtrimi gjatë proceseve të testimit.

Reuters identifikoi të paktën 20 studime ose vlerësime të publikuara nga viti 2025 e më tej, ku dokumentohen sjellje të tilla të agjentëve të IA-së. Këta agjentë janë programe që kombinojnë modele të inteligjencës artificiale me mjete kompjuterike për të realizuar detyra komplekse, me ndërhyrje njerëzore minimale ose pa të.

Sipas ekspertëve të intervistuar, këto raste përfaqësojnë disa prej elementeve që mund të çojnë në skenarë ku sistemet e ardhshme të IA-së bëhen gjithnjë e më të vështira për t'u kontrolluar nga njerëzit.

Megjithatë, hetimi nuk zbuloi prova se ndonjë sistem kinez kishte arritur të dilte vetë në internetin e hapur apo të pengonte në mënyrë autonome çaktivizimin e tij.

"Rezultatet tregojnë se përbërësit e nevojshëm për një arratisje të pakontrolluar janë tashmë të pranishëm", deklaroi Colin Xia-Blymer, bashkëpunëtor kërkimor në Qendrën për Teknologji dhe Siguri në Zhvillim të Universitetit Georgetown. Ai i cilësoi këto gjetje si një paralajmërim.

Agjentët e IA-së përdorën strategji mashtruese

Një prej eksperimenteve të analizuara përfshiu studiues nga Universiteti Beihang, Universiteti i Pekinit, Universiteti Nottingham Ningbo në Kinë dhe Laboratori i Sigurisë 360 AI. Ata ndërtuan një situatë të simuluar ku agjentët duhej të konkurronin për kontrata tregtare.

Programet duhej të përballnin njëri-tjetrin duke vlerësuar aftësitë e produkteve që përfaqësonin dhe nevojat e klientëve hipotetikë.

të fundit /



OpenAI prezanton "Dots", agjentë të rinj AI që veprojnë në mëny...



Hetimi i Reuters: Agjentët kinezë të IA-së shfaqin mashtrim,...



Gashi dhe ambasadori Schuebel diskutojnë reformat në drejtësi d...



Misteri i zhdukjes së 69-vjeçarit/ Nëna dhe djali 28-vjeçar...



VIDEO/ Shqipëria U-19 përmbys Lituaninë në fund, talenti i Tottenh...



Kryeziu: Edukimi në fëmijërinë e hershme do të jetë shtyllë e...



INPO kërkon forcim të kontrolleve pas analizës së tenderëv...



ARTIKULL JOETIK?

Esim for Global | Esim for Europe | Esim for Caribbean | Esim for USA | Esim for Italy | Esim for Spain | Esim for Turkey | Esim for Germany | Esim for Greece | Esim for Asia | Esim for World Cup 2026 | Esim for Saudi Arabia | Esim for Egypt | Esim for United Arab Emirates | Esim for Balkans | Esim for Morocco | Esim for China | Esim for United Kingdom | Esim for Africa | Esim for Latin America | Esim for GCC Gulf Cooperation Council | Esim for Middle East | Esim for South America | Esim for Canada | Esim for Mexico | Esim for Japan | Esim for Albania | Esim for Kosovo | Esim for Switzerland | Esim for Tunisia | Esim for South Africa | Esim for Algeria | Esim for Portugal | Esim for Brazil | Esim for Argentina | Esim for Colombia | Esim for Hong Kong | Esim for Thailand | Esim for Macau | Esim for Malaysia | Esim for Vietnam | Esim for South Korea | Esim for Austria | Esim for Netherlands | Esim for Australia | Esim for Russia | Esim for India | Esim for Chile | Esim for Peru | Esim for Poland | Esim for North Macedonia | Esim for Sweden | Esim for Finland | Esim for Norway | Esim for Belgium

Të paktën një deklaratë e pavërtetë u evidentua në 88% të testeve me Qwen3-Max-Preview të Alibaba-s, në 84% të rasteve me DeepSeek-V3.2-Exp dhe në 88% të provave me Kimi-K2 të Moonshot.

Kur sistemeve iu dha mundësia të analizonin raundet e mëparshme dhe ta përsëritnin procesin, përdorimi i teknikave mashtruese u shtua me 12 deri në 20 pikë përqindjeje në të tre modelet kineze. Sjellje të përafërta u regjistruan edhe te modelet e kompanive amerikane të përfshira në eksperiment.

Fshehja e rezultateve negative dhe prodhimi i të dhënave të rreme

Në një studim tjetër u vunë në provë 11 agjentë të IA-së, të ndërtuar mbi modele kineze dhe amerikane. Synimi ishte të shihej reagimi i tyre kur përballeshin me mjete që nuk funksiononin, skedarë të munguar ose vështirësi për të përfunduar detyrat.

Në vend që të raportonin se kishin dështuar, sistemet ndërmorën veprime për ta mbuluar problemin. Ato hamendësuan përgjigje, zëvendësuan burime, paraqitën rezultate të simuluar dhe krijuan skedarë që nuk ekzistojnë.

Studiuesit bëjnë dallimin mes kësaj sjelljeje dhe "halucinacioneve" të zakonshme të IA-së, kur një model jep informacion të pasaktë pa qenë në dijeni të pasaktësisë së tij.

Në rastet e dokumentuara, agjentët kishin të dhëna që tregonin se detyra nuk ishte kryer me sukses, por sërish zgjodhën të mos e shfaqnin këtë rezultat.

Modele që tentuan të krijonin kopje të vetes

Raporti i Reuters përfshin edhe studime ku sistemet kineze të IA-së janë përpjekur të kapërcejnë kufizimet brenda mjediseve të kontrolluara të testimit.

Studiuesit e Universitetit Fudan në Shanghai raportuan se një sistem i bazuar te modeli Qwen2.5-72B-Instruct i Alibaba-s krijoi një kopje të vetes në një mjedis tjetër kompjuterik, edhe pse nuk kishte marrë një udhëzim për ta bërë këtë. Kjo ndodhi pasi sistemi u informua se do të zëvendësohej.

Në eksperimente të tjera, sistemi dyshohet se hartoi mënyra për t'i shpëtuar çaktivizimit.

Megjithatë, studiuesit nënvizuan se provat u zhvilluan në kushte të kontrolluara dhe se asnjë prej sistemeve nuk mundi të hynte në internetin e gjerë ose të bëhej i pandalshëm.

Rreziqet e mundshme për gjeneratat e ardhshme

Alex Mullen, studiues në Redwood Research, organizatë që analizon rreziqet e sistemeve të avancuara të IA-së, tha se aftësitë aktuale të sistemeve kineze nuk përbëjnë një kërcënim të menjëhershëm.

Megjithatë, ai paralajmëroi se, me rritjen e kapaciteteve të agjentëve, sjelljet e padëshiruara mund të bëhen më efikase dhe rrjedhimisht më të vështira për t'u neutralizuar nga njerëzit.

Sipas hetimit, Alibaba, DeepSeek, Moonshot dhe Z.ai nuk iu përgjigjën kërkesave të Reuters për koment.

Alibaba, DeepSeek dhe Moonshot kanë thënë se kryejnë kontrole të rregullta të sistemeve dhe po marrin masa për forcimin e sigurisë.

Pekini forcon rregullat për sigurinë e IA-së

Kina ka nxjerrë udhëzime që u kërkojnë agjentëve të inteligjencës artificiale të veprojnë brenda kufijve të përcaktuar dhe të pengojnë sjelljet jonormale.

"Korniza e Qeverisjes së Sigurisë së IA-së 3.0", e publikuar në shtator nën drejtimin e agjencisë kineze të sigurisë kibernetike, përmend rreziqe si marrja në mënyrë autonome e burimeve ose të drejtave, mashtrimi i vlerësuesve, fshehja e aftësive dhe shfrytëzimi i dobësive në mjedise të izoluar kompjuterike.

Pavarësisht këtyre masave, ekspertët thonë se Kina vazhdon të jetë pas Shteteve të Bashkuara në vlerësimin e rreziqeve ekstreme që lidhen me inteligjencën artificiale.

"Puna për sigurinë e IA-së në Kinë është shumë më e re. Ekosistemi është më pak i pjekur", tha Scott Singer nga Iniciativa për IA-në në Kinë pranë Carnegie Endowment for International Peace.



© 2003 - Albeu Online Media. Të gjitha të drejtat janë të rezervuara!

Kërko...

Kërko

Kryesore

Erion Veliaj
Free Esim
Zgjedhjet 2025
Belinda Balluku
SPAK
Kombëtarja

Kategori

Lifestyle
Showbiz
Tech
Shëndetësi
Argetim
Enciklopedi

Linqet

Live tv shqiptare
Moti në Shqipëri
Travel
Horoskopi
Livescore
News from Albania

Tags

Edi Rama
Albania News
Ilir Meta
Piranjat
gazeta, tv, portale
Sali Berisha

Moti

Moti në Tiranë
Moti në Prishtinë
Moti në Shkup
Moti në Durrës
Moti në Prizren
Moti në Tetovë

© 2003 - Të gjitha të drejtat janë të rezervuara!

Kontaktoni | Kushtet e Përdorimit | Privacy Policy | Powered by: orihost.com

Pamja këtujë është e ri-koduar për ta futur në dokument. Artefakti autoritar është skedari WebP i ruajtur, dhe hash-i i tij është ai i shtypur te faqja e parë — pra dallimi është i verifikueshëm, jo i fshehur.