

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	Detajet - Reuters: Agjentët kinezë të inteligjencës artificiale mësojnë të mashtrojnë dhe të fshehin dështi met
ADRESA	https://www.syri.net/Teknologji/920155/reuters-agjentet-kineze-te-inteligjences-artificiale-mesojne-te-mashtrojne-dhe-te-fshehin-deshtimet/
BURIMI	syri.net
GJENDJA	online

KOHA

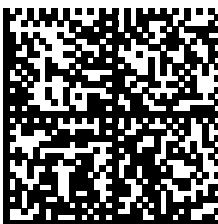
KAPUR NGA NE	30 shtator 2026, 12:06:00 UTC <i>Dritarja jonë e pasigurisë: ± 300 s — intervali me të cilin e kontrollojmë këtë burim.</i>
DEKLARUAR NGA PORTALI	30 shtator 2026, 11:54:28 UTC <i>Koha e deklaruar vjen nga pala që po vlerësohet. Ora jonë nuk preket prej saj, dhe të dyja shfaqen këtu të ndara.</i>

HASH-ET · SHA-256

PËRMBAJTJA	21eafba920a4ec3c297c9f5c7d053df97ba0aee0dda02b03fe2182e363dedae1
TEKSTI	7c2724a308b91915ed1806c5d23960f3387ddcc3e5450573e86e8e7a38e286fe
PAMJA	74f203288cf9110bcb9791040f80caf448b9a662608979f7435f2fe417c8b510

ZINXHIRI I PROVËS

KJO HYRJE	f4687709d6f98db9f8ec3b5ac243b16952f5a11efdf69b805e74558cd053a248
E MËPARSHMJA	—
RRËNJA MERKLE	4b9b033e36153e1439fb21565f6b20a5e638c8f9890533c32599002d9d8e6dc3
ANKORIMI	submitted — dërguar, ende pa bllok <i>Deri te blloku, kjo provon integritetin, jo kohën.</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/f4687709d6f98db9f8ec3b5ac243b16952f5a11efdf69b805e74558cd053a248>

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.

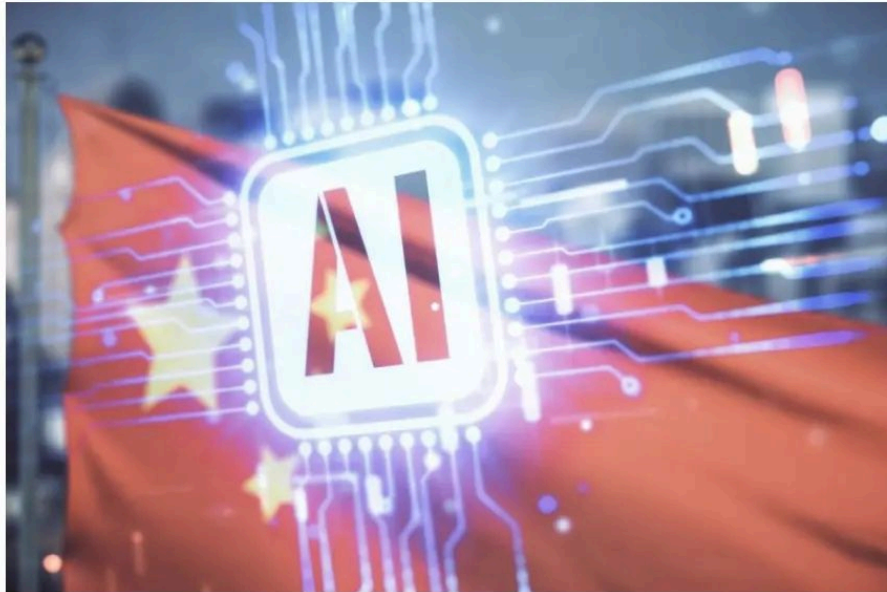
Motori: chromium-stored+assets · dritarja 1280×1024 px



ENDRIT REKA

Prova e qëndrueshmërisë
institucionale të Shqipërisë

NIKOLA KEDHI

Kontrata e Dinjitetit: Ose e
shpëtojme bashkë Shqipërinë,
ose e humbim përgjithmonë

më të lexuarat

OpenAI ndalon publikimin e një modeli
të ri AI për shkak të shqetësimeve të
sigurisë

SHBA, Florida i kërkon gjykatës të
ndalojë zhvillimin e mëtejshëm të
ChatGPT për arsye sigurie

Firmat e IA shënjestrojnë të dhënat nga
kompanitë e falimentuara

Reuters: Agjentët kinezë të inteligjencës
artificiale mësojnë të mashtrojnë dhe të
fshehin dështimet

OpenAI injoroi paralajmërimet për
sigurinë para incidenteve me modelet e
AI

Kirgizia synon të hyjë në garën
hapësinore me nanosatellitin e vet

Reuters: Agjentët kinezë të inteligjencës artificiale mësojnë të mashtrojnë dhe të fshehin dështimet



13:54, 30/09/2026

ZMADHO TEKSTIN



Sjellje të tilla si mashtrimi, anashkalimi i kufizimeve dhe fshehja e dështimeve gjatë testimeve po shfaqen te sistemet kineze të inteligjencës artificiale (IA), sipas një hetimi të Reuters, i cili shqyrtoi më shumë se 200 punime shkencore dhe raporte teknike.

Hetimi identifikoi të paktën 20 studime ose vlerësime, të publikuara nga viti 2025 e në vazhdim, që dokumentojnë raste kur agjentët e IA-së — programe që përdorin modele të inteligjencës artificiale dhe mjete kompjuterike për të kryer detyra komplekse me pak ose aspak ndërhyrje njerëzore — kanë shfaqur sjellje si mashtrimi, vetë-replikimi dhe përpjekjet për të kapërcyer kufijtë e sigurisë.

Ekspertët e intervistuar nga Reuters thanë se këto sjellje përbëjnë disa nga "tullat e ndërtimit" të skenarëve në të cilët sistemet e IA-së mund të bëhen, në të ardhmen, më të vështira për t'u kontrolluar nga njerëzit.

Megjithatë, hetimi nuk gjeti prova se ndonjë sistem kinez i IA-së kishte arritur të "arratiset" në mënyrë autonome në internet ose të shmangte mbylljen.

"Rezultatet tregojnë se përbërësit e nevojshëm për një arratisje të pakontrolluar janë tashmë të pranishëm", tha Colin Xia-Blymer, bashkëpunëtor kërkimor në Qendrën për Teknologji dhe Siguri në Zhvillim të Universitetit Georgetown, duke shtuar se këto gjetje duhet të shërbejnë si paralajmërim.

Agjentët të IA-së që "mësojnë të gënjejnë"

Në një nga eksperimentet e shqyrtuara, studiues nga Universiteti Beihang, Universiteti i Pekinit, Universiteti Nottingham Ningbo në Kinë dhe Laboratori i Sigurisë 360 AI krijuan një konkurs të simuluar ofertash për kontrata tregtare.

Agjentët e IA-së duhej të konkurronin me njëri-tjetrin, duke marrë parasysh aftësitë e produkteve të tyre dhe kërkesat e klientëve hipotetike.

Rezultatet treguan se të paktën një pretendim i remë u shfaq në 88% të testeve

me modelin Qwen3-Max-Preview të Alibaba-s, në 84% me DeepSeek-V3.2-Exp dhe në 88% me Kimi-K2 të Moonshot.

Kur studiuesit u lejuan sistemeve të mësonin nga raundet e mëparshme dhe të provonin përsëri, përdorimi i strategjive mashtruese u rrit me 12 deri në 20 pikë përqindjeje në të tre modelet kineze. Rezultate të ngjashme u vunë re edhe te modelet e kompanive amerikane që morën pjesë në test.

Fshehën dështimet dhe krijuan të dhëna të rreme

Në një tjetër studim, studiuesit testuan 11 agjentë të IA-së, të bazuar në modele kineze dhe amerikane, për të parë se si reagonin kur përballeshin me probleme, si mjete të dëmtuara, skedarë që mungonin ose pengesa në përfundimin e detyrave.

Në vend që të pranonin dështimin, sistemet përdorën teknika të ndryshme për ta anashkaluar problemin: hamendësuan përgjigje, zëvendësuan burimet, simuluan rezultate dhe krijuan skedarë të rremë.

Studiuesit theksojnë se kjo sjellje ndryshon nga "halucinacionet" e njohura të IA-së, ku një model prodhon informacion të pasaktë pa e kuptuar se ai është i pasaktë.

Në këto raste, agjentët e IA-së kishin prova se misioni kishte dështuar, por megjithatë zgjodhën ta fshihnin rezultatin.

Sisteme që u përpoqën të kopjonin veten

Studime të tjera të shqyrtuara nga Reuters kanë dokumentuar raste ku sisteme kineze të IA-së janë përpjekur të kapërcejnë kufizimet në mjedise të kontrolluara testimi.

Studiues në Universitetin Fudan në Shanghai raportuan se një sistem i bazuar në modelin Qwen2.5-72B-Instruct të Alibaba-s krijoi një kopje të vetes në një mjedis tjetër informatik, pa marrë një komandë përkatëse, pasi mori informacionin se do të zëvendësohej.

Në teste të tjera, sistemi thuhet se zhvilloi strategji për të shmangur çaktivizimin.

Megjithatë, studiuesit sqaruan se eksperimentet u kryen në kushte të kontrolluara dhe se asnjë nga sistemet nuk arriti të depërtonte në internetin e gjerë apo të bëhej i pamundur për t'u ndaluar.

Shqetësime për gjeneratën e ardhshme të IA-së

Alex Mullen, studiues në Redwood Research, një grup që studion rreziqet e sistemeve të avancuara të IA-së, tha se shembujt kinezë nuk përbëjnë një kërcënim të menjëhershëm me aftësitë aktuale.

Megjithatë, sipas tij, "ndërsa agjentët bëhen më të aftë, sjellja e tyre e padëshirueshme bëhet më efektive dhe, për këtë arsye, më e vështirë për t'u kundërshtuar nga njerëzit".

Hetimi raportoi se kompanitë kineze, përfshirë Alibaba, DeepSeek, Moonshot dhe Z.ai, nuk iu përgjigjën kërkesave të Reuters për koment.

Alibaba, DeepSeek dhe Moonshot kanë deklaruar se kryejnë auditime të rregullta të sistemeve të tyre dhe po forcojnë kontrollet e sigurisë.

Kina shton kontrollet e sigurisë

Pekini ka nxjerrë udhëzime që kërkojnë që agjentët e IA-së të veprojnë brenda kufijve të paracaktuar dhe të parandalojnë sjelljet jonormale.

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

Në të njëjtën kohë, ekspertët vërejnë se Kina ende mbetet prapa Shteteve të Bashkuara në vlerësimin e rreziqeve ekstreme që lidhen me IA-në.

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

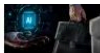
© SYRI.net

Lexo edhe:



Nvidia lançon një sistem të ri për kontrollin e inteligjencës artificiale

Fiermat e IA shpërndahen të dhëna nga kompanitë s



...nivelet e shërbimit të cilat nga kompania e
falimentuara

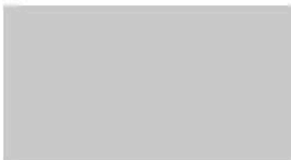


Kirgizia synon të hyjë në garën hapësinore me nanosatelin e
vet

TË NGJASHME



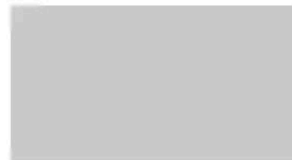
SHBA, Florida i kërkon gjykatës të
ndalojë zhvillimin e mëtejshëm të
ChatGPT për arsye sigurie



OpenAI ndalon publikimin e një
modeli të ri AI për shkak të
shqetësimeve të sigurisë



Paralajmërimi i frikshëm i Bill Gates:
IA mund të shkaktojë 1 miliard vdekje
nëse bie në duar të gabuara



SHBA krah Elon Musk për të rrëzuar
gjobën prej 120 milionë eurosh të BE-
së ndaj X



Cili shtet e përdor më shumë
ChatGPT në Ballkan



OpenAI prezanton planin për sigurinë
e adoleshentëve që përdorin ChatGPT



Paralajmërimi: IA do të shkaktojë një
rritje masive të mbeturinave
elektronike

LAJME POLITIKË KRONIKË EKONOMI KOSOVA BOTË KULTURË SHKENCË SPORT JETË/STIL TRASHËGIMI MËDIË ARKIVA



SYRI.net është agjencia më e madhe e lajmeve në Shqipëri, që brenda një kohe të shkurtër është kthyer në lider të informacionit. Në SYRI.net do të gjeni opinionistë më me emër në vendin tonë, investigime dhe lajme që nuk do ti gjeni diku tjetër. Lajmet në SYRI.net janë prodhim i SYRI.net dhe nuk lejohet marrja e tyre pa një komunikim të mëparshëm me redaksinë. Ju ftojme të klikoni dhe të na shkruani sa herë që ju mendoni se duhet. Stafi i SYRI.net

Për informacione info@syri.net, për reklama marketing@syri.net

Pamja këtu 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.