

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	"Agjentët kinezë të inteligjencës artificiale mësojnë të mashtrojnë dhe të fshehin dështimet"
ADRESA	https://lajmifundit.al/agjentet-kineze-te-inteligjences-artificiale-mesojne-te-mashtrojne-e-dhe-te-fshehin-deshtimet-i1339242
BURIMI	lajmifundit.al
GJENDJA	online

KOHA

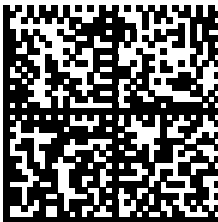
KAPUR NGA NE	30 shtator 2026, 12:06:05 UTC <i>Dritarja jonë e pasigurisë: ± 900 s — intervali me të cilin e kontrollojmë këtë burim.</i>
DEKLARUAR NGA PORTALI	30 shtator 2026, 13:54:42 UTC <i>Koha e deklaruar vjen nga pala që po vlerësohet. Ora jonë nuk preket prej saj, dhe të dyja shfaqen këtu të ndara.</i>

HASH-ET · SHA-256

PËRMBAJTJA	aa61424cd5ff64460b60d24e82baff06345a6ac8147039661ce53c7a26e86b7d
TEKSTI	fd5f43c0f5baa0eb484a030f27865f8e379aa0ed316999fd4b578dbcc2049253
PAMJA	8350e9d6b636ee5d8d51c3e684ee8a0ffeade4adf14d68fbbd3faa36cbe4b8da

ZINXHIRI I PROVËS

KJO HYRJE	576ee5e28404b9ebac8ed062ffb13a3a7c10a6687598cb3b420a961148e852c6
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/576ee5e28404b9ebac8ed062ffb13a3a7c10a6687598cb3b420a961148e852c6>

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

"Agentët kinezë të inteligjencës artificiale mësojnë të mashtrojnë dhe të fshehin dështimet"

[Lexo lajmin e plotë](#) →

Profesionet që po tërheqin gjithnjë e më shumë femra, ofrojnë mundësi për karrierë të suksesshme

[Lexo lajmin e plotë](#) →

Aksidenti në kantierin detar në Ankona, humb jetën punëtori, pl'agoset rëndë një tjetër

[Lexo lajmin e plotë](#) →

Tragjedi në metro! Çanta i ngeci në dyer, 57-vjeçari përfundon në shina dhe humb jetën

[Lexo lajmin e plotë](#) →

Mbi 140 kg, zbulohen detaje për shëndetin e liderit të Koresë së Veriut, Kim Jong-un

[Lexo lajmin e plotë](#) →

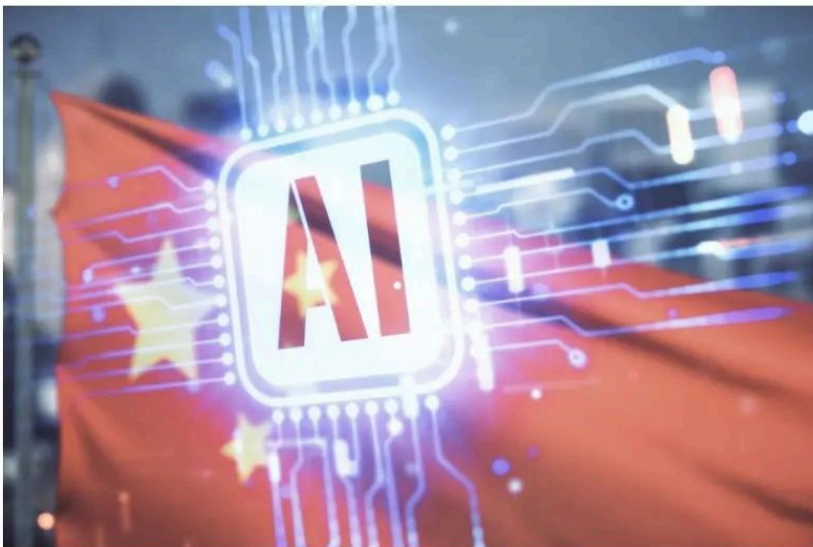
"Agentët kinezë të inteligjencës artificiale mësojnë të mashtrojnë dhe të fshehin dështimet"



LAJMI FUNDIT / 30 SEPTEMBER 2026, 13:54

Shpërndaje

BOTA



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 agentët 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 "arratisë" 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.

MOTI NE TIRANE

Tirana, Albania

POLITIKË

Detajet, nga burgu në Bashkinë e Tiranës, Veliq firmos ndryshimet e reja në kryeqytet

POLITIKË

Situata politike në vend, Shehi: Nëse vazhdon ky sistem dhe kjo mënyrë qeverisjeje, do të kalcifikohet e keqja!

TRENDING



Domosdoshmëria e sigurimit të aseteve fizike në pronësi të shtetit



Destinacioni sekret: Wizz Air sjell në TIA 150 influencers dhe gazetarë për të zbuluar Shqipërinë

LEXONI GJITHASHTU



Aksident në kantierin detar në Ankona, humb jetën punëtori, pl'agoset rëndë një tjetër



Mbi 140 kg, zbulohen detaje për shëndetin e liderit të Koresë së Veriut, Kim Jong-un



Si i kapi FBI-ja dy vëllezërit shqiptarë? Drejtonin tregun e identiteteve të vjedhura

Agjentë 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 konkuronin me njëri-tjetrin, duke marrë parasysht aftësitë e produkteve të tyre dhe kërkesat e klientëve hipotetikë.

Rezultatet treguan se të paktën një pretendim i rremë 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 praninin 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 zgjedhën 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 Shangai 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.

POLITIKË

Ish-gjenerali që u bë ambasador/ Eric Wendt nis "betejën" diplomatike amerikane, zbulohet cilin takon të parin të mërkurën

POLITIKË

Këshilli Bashkiak i Tiranës miraton propozimin e Veliajt, 189 zjarrëfikës përfitojnë një pagë shpërblim

ME TE LEXUARAT



"Zbrazti karika'torin ndaj shqiptarëve që i hynë në shtëpi" – Mediat italiane: Jeta e "dyfishtë" e...

[Lexo lajmin e plotë](#)



Euron pëson rënien si asnjëherë më parë, me sa tregtohet sot, këmbimi valutor 23 shtator 2026

[Lexo lajmin e plotë](#)



Skandali "Xhaka" merr përmasa të tjera, mjekja del kundër kapitenit: E kam vaksinuar vetë

[Lexo lajmin e plotë](#)



Zbulohet e vërteta për Bleona Matën, avokati: E ëma në gjendje të rëndë! A ndodhet në Greqi?!

[Lexo lajmin e plotë](#)

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.

ME TE LEXUARAT

“Zbrazì karika'torin ndaj shqiptarëve që i hynë në shtëpi” – Mediat italiane: Jeta e “dyfishtë” e Gjovalin Lazrit

[Lexo lajmin e plote](#)

NA SHKRUAJ

Portali Lajmifundit sjell Lajme nga Shqipëria, rajoni dhe bota në kohë rekord. Ju lutemi na kontaktoni nëse doni të kontribuoni me informacione të vlefshme për publikim në portalin tonë.

[Kliko për më shumë informacion](#)

Euron pëson rënien si asnjëherë më parë, me sa tregtohet sot, këmbimi valutor 23 shtator 2026

[Lexo lajmin e plote](#)

Skandali “Xhaka” merr përmasa të tjera, mjekja del kundër kapitenit: E kam vaksinuar vetë

[Lexo lajmin e plote](#)

Zbulohet e vërteta për Bleona Matën, avokati: E ëma në gjendje të rëndë! A ndodhet në Greqi?!

[Lexo lajmin e plote](#)

"Askush nuk është mbi ligjin"/
Rama shkarkon menjëherë Vlorë
Hysenin: Pas kthimit në Tiranë
zgjedh drejtuesin e ri të SHISH

[Lexo lajmin e plotë](#)

Sekuestrohet hoteli "Central Park
Premium" në Ksamil, u ndërtua pa
leje! Ja kujt i përket

[Lexo lajmin e plotë](#)

"Ka gjëra që edhe ne që i dimë nuk
flasim dot"/ Artan Hoxha: Ka fakte
tronditëse për Ballukun, gropë u
hap

[Lexo lajmin e plotë](#)

Lajmi Fundit



KUSH JEMI

KONTAKT REDAKSINË

TE DHENAT PERSONALE

COOKIES

LajmiFundit merr pjesë në programe të ndryshme marketingu ku mund të përfitojë komisione për produktet e zgjedhura në artikujt e publikuar.

Created & Monetized by **Mediadesk**

REGJISTROHU NË NEWSLETTER

OK

Ju do të mëmë njoftime periodike me Lajmet e Fundit.

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.