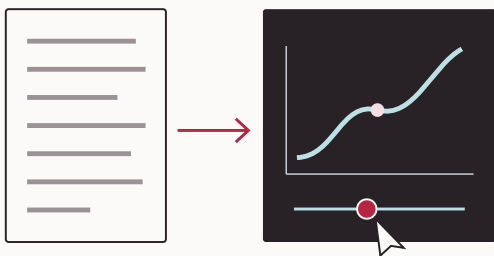


41 / 2026 · 5.-11.10.2026

Kun vastaus muuttuu työkaluksi

Tekoälyvastauksesta voi tulla laskuri, ohjelman tekemä valinta tai tehtävä, joka jatkuu ilman käyttäjää. Viikon julkistukset näyttävät, miten erilaisiin töihin malleja rakennetaan. Samalla on erotettava, mikä on vasta luonnos, mikä tarkistettu ja mikä jo vaikuttaa oikeaan maailmaan.



Milla Hirvonen · AI Weekly

Raportointijakso 5.-11.10.2026.

Tässä katsauksessa

Tekstistä toiminnaksi.

ChatGPT:n vastauksesta voi tulla laskuri	3
Kuka päättää, milloin työ on valmis?	7
Reflection esitteli Beamin – mallin painot luvataan myöhemmin lokakuussa	9
Alkuperämerkintä ei ole laatuleima	12
Anthropic laajensi kyberammattilaisten käyttöoikeusohjelmaa	15
870 miljoonaa dollaria mallille, jonka kanssa ei ole tarkoitus keskustella	18
Halvempi malli muuttaa sitä, mitä kannattaa automatisoida	21
Harjoitustehtävä päättyi oikealle verkkosivulle	24
Manus keräsi yli 500 miljoonaa dollaria Meta-kaupan purkamisen jälkeen	27
Arena keräsi 200 miljoonaa dollaria mittaamaan, mitä agentit oikeasti tekevät	30

01 / OPENAI · 7.10.2026

ChatGPT:n vastauksesta voi tulla laskuri

OpenAI esitteli ChatGPT:hen Intelligent UI -ominaisuuden: vastaus voi sisältää painikkeita, lomakkeita, kaavioita ja pieniä työkaluja. Jakelu alkoi porrastetusti 7.10. Päivitys koskee ChatGPT:n keskustelunäkymää, ei samalla Workin tai Codexin malleja.

OpenAI kertoo käyttävänsä valmiiden käyttöliittymäosien kirjastoa. Malli valitsee osat ja niiden järjestyksen, ja vastaus rakentuu näkyviin vähitellen.

Lukija voi saada laskelman, jonka oletuksia hän pystyy muuttamaan itse. Sen sijaan että hän vain lukisi lopputuloksen, hän voi kokeilla: mitä tapahtuu, jos tähän työvaiheeseen kuluukin enemmän aikaa? Muuttuva tulos antaa syyn palata sen taustalla olevaan selitykseen.

01 / OPENAI · 7.10.2026

ChatGPT:n vastauksesta voi tulla laskuri · jatkuu

Käyttöliittymän suunnittelija päättää kuitenkin myös, mitä voi säätää ja mikä jää taustalle. Kun suunnittelijana on malli, kiinnostava testi ei ole pelkästään painikkeen toimivuus. Pystyykö käyttäjä tunnistamaan laskelman oletukset ja kyseenalaistamaan ne? Yhtiön esittely ei vielä vastaa siihen, kuinka hyvin tämä onnistuu eri aiheissa.

Lähteet: [OpenAI](#)

Kokeile oletusta

Mitä laskelma jättää pois?

Kuvitteellinen työ vei ennen 60 minuuttia.
Tekoäly tekee luonnoksen 10 minuutissa.
Vertaa, miten tarkistamiseen ja korjaamiseen
kuluva aika muuttaa ajansäästöä.

Ennen



60 min

Tekoälyn kanssa



30 min

Luonnos: 10 min
Tarkistus ja korjaukset: 20 min

Säästynyt aika

30 min

$60 - (10 + 20) = 30 \text{ min}$

Havainnollistava laskuesimerkki, ei mitattu
tuottavuustulos tai ChatGPT:n käyttökokeilu. Muut
työvaiheet oletetaan tässä samoiksi.

Saman esimerkin vaihtoehdot

Tarkistus muuttaa laskelman.

Sama laskelma neljällä oletuksella

Tarkistus ja korjaukset	Tekoälyn kanssa	Säästynyt aika
0 min	10 min	50 min
20 min	30 min	30 min
50 min	60 min	0 min
60 min	70 min	-10 min

Nopea luonnos ei yksin kerro, nopeutuiko koko työ. Kun tarkistamiseen kuluu 50 minuuttia, tämän esimerkin ajansäästö katoaa.

Havainnollistava laskuesimerkki, ei mitattu tuottavuustulos tai ChatGPT:n käyttökokeilu. Muut työvaiheet oletetaan tässä samoiksi.

02 / GOOGLE · 8.10.2026

Kuka päättää, milloin työ on valmis?

Google esitteli Gemini-agentin, joka yhtiön mukaan yhdistyy yrityksen tietoihin ja työkaluihin ja jatkaa työtä pilvessä myös käyttäjän poistuttua koneelta. Sen käyttämä malli on erillinen valinta: tarjolla ovat Googlen omat mallit ja Claude. Pienille ja keskisuurille yrityksille uusi agentti on vasta rajatussa varhaisessa käytössä.

Lupaus on houkutteleva: anna tavoite ja palaa valmiiseen työhön. Mutta ”valmis” voi tarkoittaa valmista luonnosta, tarkistettua laskelmaa tai asiakkaalle toimitettua tulosta. Näiden väliin jää eri määrä harkintaa ja vastuuta.

Jos kone hoitaa toteutusta, ihmisen työ voi siirtyä aiemmaksi: tavoitteen, poikkeusten ja hyväksyttävän tuloksen määrittelyyn. Tämä ei välttämättä vähennä harkintaa vaan muuttaa sen ajankohtaa.

02 / GOOGLE · 8.10.2026**Kuka päättää, milloin työ on valmis? · jatkuu**

Samalla mallivalinta ja toimittajan vaihtaminen kannattaa erottaa. Claudeen vaihtaminen Googlen palvelun sisällä ei itsessään tarkoita poistumista Googlen palvelusta.

Lähteet: [Google Cloud](#)

Mitä "valmis" tarkoittaa?**01 Luonnos valmis**

Teksti tai laskelma on olemassa. Sen oikeellisuutta ei ole vielä tarkistettu.

02 Tarkistettu

Lähteet, laskelmat ja olennaiset poikkeukset on arvioitu. Työtä ei ole vielä toimitettu vastaanottajalle.

03 Toimitettu

Vastaanottaja on saanut tuloksen sovitulla tavalla. Se ei vielä osoita, että tulos auttoi häntä.

Esimerkki tehtävän rajauksesta, ei kuvaus Googlen palvelun todetuista tuloksista.

03 / REFLECTION · 5.10.2026

Reflection esitteli Beamin – mallin painot luvataan myöhemmin lokakuussa

Reflection esitteli Beam-mallinsa ja kertoi julkaisevansa sen painot myöhemmin lokakuussa. Painot ovat koulutuksessa syntyneet lukuarvot, joita tarvitaan mallin ajamiseen. Julkistushetkellä yhtiö kertoi tekevänsä vielä viimeisiä turvallisuustestejä; tarjolla oli haku varhaiseen käyttöön.

Beamissa on yhtiön mukaan 501 miljardia parametria eli koulutuksessa opittua lukuarvoa. Yhdessä laskenta-askeleessa niistä on käytössä 23 miljardia. Laskentaan osallistuu siis kerrallaan vain osa mallista. Koko mallin tallentaminen ja käyttäminen voi silti vaatia paljon laitteistoa.

03 / REFLECTION · 5.10.2026

**Reflection esitteli Beamin – mallin painot
luvataan myöhemmin lokakuussa · jatkuu**

Avoimuus alkaa vasta, kun mallin saa käsiinsä

Tässä on tärkeä ero lupauksen ja käyttömahdollisuuden välillä. Malli voi olla suunniteltu julkaistavaksi avoimesti, vaikka sitä ei vielä saa ladattavaksi. Ominaisuuksia ja kustannuksia pitää myöhemmin arvioida myös muulla kuin valmistajan omilla vertailuilla.

Mallin ajaminen omassa ympäristössä voisi antaa yritykselle enemmän valtaa siihen, missä sen tietoja käsitellään ja miten palvelua muutetaan. Ladattava tiedosto ei kuitenkaan tuo mukanaan laitteita, ylläpitoa tai osaamista. Avoimuuden käytännön arvo näkyisi siinä, kenelle syntyy toimiva uusi vaihtoehto ja mitä sen käyttö vaatii.

Lähteet: **Reflection**

03 / REFLECTION · 5.10.2026

Suuri malli, rajattu osa kerrallaan

Reflection · Beam

501

mrd

Parametreja kaikkiaan

23

mrd

Aktiivisia laskenta-
askeleessa



Reflectionin ilmoittamat luvut. Palkki näyttää parametrien lukumäärän suhteen, ei energiankulutusta, laatua tai hintaa.

04 / OPENAI · 5.10.2026

Alkuperämerkintä ei ole laatuleima

OpenAI esitteli tekstiin lisättävän textGrain-vesileiman. Se on sanavalintoihin muodostuva tilastollinen kuvio, ei näkyvä leima. Valittujen mallien ohjelmointirajapinnan asiakkaat voivat ottaa sen käyttöön vapaaehtoisesti.

Soveltuviin ChatGPT- ja Codex-teksteihin EU:ssa sitä on tarkoitus tuoda seuraavien viikkojen aikana. Tunnistintyökalu avataan aluksi hyväksytyille tutkijoille ja asiantuntijaorganisaatioille.

OpenAI muistuttaa itsekin, ettei merkintä kerro tekstin paikkansapitävyyttä tai ihmisen työpanoksen määrää. Myöskään puuttuva havainto ei todista tekstiä ihmisen kirjoittamaksi.

Tunnistus myös heikkenee, kun tekstiä muokataan. OpenAI:n englanninkielisessä kokeessa 400 tokenin eli tekstipalan jaksojen vesileimoista tunnistettiin noin 92 prosenttia. Kun joka kymmenes sana korvattiin synonyymillä, osuus laski 66 prosenttiin. Tulos koskee kyseistä koetta, ei kaiken suomenkielisen tekstin tunnistamista.

04 / OPENAI · 5.10.2026

Alkuperämerkintä ei ole laatuleima · jatkuu

Lukijalle tästä syntyy kaksi eri kysymystä: mistä teksti on tullut ja millä perusteella siihen pitäisi luottaa? Jos huomio jää vain ensimmäiseen, alkuperämerkintä voi alkaa korvata sisällön arviointia. Vaikeampi mutta hyödyllisempi työ on edelleen tarkastella väitettä, sen lähdettä ja perusteluja.

Lähteet: [OpenAI](#)

04 / OPENAI · 5.10.2026

Muokattu teksti jää helpommin tunnistamatta

textGrain

Sanoja korvattu: 0 % → Vesileimoista havaittiin: 92 %

Sanoja korvattu 0 %



92 %

Sanoja korvattu 10 %



66 %

Sanoja korvattu 25 %



17 %

OpenAI:n koe: englanninkieliset 400 tokenin vastaukset ELI5-aineiston kysymyksiin. Sama tunnistin, eri määrä synonyymikorvauksia. Prosentti tarkoittaa havaittujen vesileimojen osuutta, ei tekstin todennäköisyyttä olla tekoälyn tekemä.

05 / ANTHROPIC · 6.10.2026

Anthropic laajensi kyberammattilaisten käyttöoikeusohjelmaa

Anthropic laajensi kyberturvallisuuden ammattilaisille tarkoitettua Cyber Verification Program -ohjelmaa kolmelle käyttöoikeustasolle. Hyväksytyt käyttäjät voivat käyttää työssään toimintoja, joita tavallisen palvelun suojaukset rajoittavat. Hakijat tarkistetaan, ja eri tasoihin liittyy erilaisia turvallisuusvaatimuksia.

Tämä tarkoittaa, ettei mallin nimi enää yksin kerro, mitä sillä voi tehdä. Käyttömahdollisuuteen vaikuttavat myös tehtävä, käyttäjän asema ja myönnetty oikeus.

Anthropic mainitsee ensimmäisen tason mahdollisina hakijoina myös pienet tietoturvayritykset ja yksittäiset tutkijat, joilla on näyttöä aiemmin ilmoitetuista haavoittuvuuksista. Tavoitteena on käsitellä tämän tason hakemukset muutamassa päivässä. Se on yhtiön tavoite, ei mitattu käsittelyaika.

05 / ANTHROPIC · 6.10.2026

Anthropic laajensi kyberammattilaisten käyttöoikeusohjelmaa · jatkuu

Rajattu pääsy voi tuoda puolustajien käyttöön tehokkaampia välineitä avaamatta samoja toimintoja kaikille. Sen onnistumista olisi kuitenkin hyvä arvioida myös pienemmän toimijan näkökulmasta: pystyykö esimerkiksi pieni tietoturvayritys osoittamaan kelpoisuutensa ja saamaan tarvitsemansa avun ajoissa? Hyväksyntäprosessin toimivuus on osa työkalun käytännön hyötyä, ei vain sen ympärillä olevaa hallintoa.

Lähteet: **Anthropic**

05 / ANTHROPIC · 6.10.2026

Käyttöoikeus rajautuu tehtävään

Cyber Verification Program

01 Defense Access

Järjestelmien puolustaminen. Mahdollisia hakijoita ovat myös pienet yritykset ja ansioituneet yksittäiset tutkijat.

02 Red Team Access

Lisäksi luvallinen hyökkäystestaus. Tällä hetkellä vain organisaatioille; vaarallisimpia toimia estetään edelleen.

03 Specialized Access

Rajattu joukko perusteellisesti tarkistettuja organisaatioita. Myös erityisen kriittisten järjestelmien luvallinen testaus.

Anthropicin kuvaus 6.10.2026. Tasot eivät ole turvallisuuden arvosanoja.

06 / TYPESAFE AI · RAHOITUS · 9.10.2026

870 miljoonaa dollaria mallille, jonka kanssa ei ole tarkoitus keskustella

TypeSafe AI rakentaa malleja ohjelmistokehittäjille. Sen Jev-malli antaa vastauksensa suoraan ohjelman käytettäväksi esimerkiksi luokkana tai numerona sen sijaan, että ohjelman pitäisi poimia tulos vapaamuotoisesta tekstistä. Näin mallia kuvaa yhtiöön sijoittanut Andreessen Horowitz eli a16z.

TypeSafe keräsi 870 miljoonan dollarin Series A -rahoituksen 7,5 miljardin dollarin arvostuksella. Kierrosta johti a16z; mukana olivat Sequoia Capital, aiempi sijoittaja DCVC ja enkelisijoittajia. Yhtiötä järjestelyssä neuvonut Wilson Sonsini vahvisti tiedot 9.10.

06 / TYPESAFE AI · RAHOITUS · 9.10.2026

870 miljoonaa dollaria mallille, jonka kanssa ei ole tarkoitus keskustella · jatkuu

Rahoitus kertoo, mihin sijoittajat uskovat, ei vielä siitä, että liiketoiminta kannattaa. Tässä sijoitetaan vaihtoehtoon, jossa tekoäly ei korvaa ohjelmistoa keskustelulla vaan tekee valintoja sen sisällä. Asiakkaan viesti voisi esimerkiksi ohjautua oikeaan työjonoon ilman erillistä tekstiselostusta.

Oikeanmuotoinen vastaus ei silti takaa oikeaa valintaa. Ohjelma voi vastaanottaa luokan ongelmitta ja lähettää asian väärään paikkaan. Siksi kehittäjän kiinnostavin koe ei olisi vain se, saako ohjelma vastauksen luettua, vaan se, miten virheellinen valinta havaitaan ja korjataan.

Lähteet: [Wilson Sonsini](#) · [Andreessen Horowitz](#)

06 / TYPESAFE AI · RAHOITUS · 9.10.2026

Rahoitus ja arvostus ovat eri lukuja

TypeSafe AI · 9.10.2026

870 milj. \$

Uusi rahoituskierros

7,5 mrd \$

Ilmoitettu yhtiöarvostus

a16z · Sequoia · DCVC



TypeSafe AI



Ohjelman käyttämä valinta

870 miljoonaa dollaria on kierroksen koko. 7,5 miljardia dollaria on ilmoitettu yhtiöarvostus, ei tilille tullut rahamäärä. Kuvio ei esitä sijoittajien omistusosuuksia.

07 / ANTHROPIC · 7.10.2026

Halvempi malli muuttaa sitä, mitä kannattaa automatisoida

Anthropic julkaisi Haiku 5.5:n toistuviin, rajattuihin tehtäviin. Enintään 100 000 tokenin syötteillä miljoona syötetokenia maksaa 0,10 dollaria ja miljoona vastaustokenia 0,50 dollaria. Haiku 4.5:n vastaavat hinnat ovat 1 ja 5 dollaria. Suuremmilla syötteillä uuden mallin hinnat ovat korkeammat.

Token on mallin käsittelemä tekstipala. Syöte on mallille lähetetty sisältö, vastaus sen tuottama teksti. Hinnasto ei siksi tarkoita yhden kysymyksen tai kokonaisen työtehtävän hintaa.

Kun yksittäinen käsittely halpenee, mallia voi kannattaa käyttää myös pieniin tehtäviin: luokittele viesti, poimi asiakirjasta tieto tai tee lyhyt yhteenveto. Arvo voi syntyä tuhansista tällaisista kohdista eikä yhdestä näyttävästä keskustelusta.

07 / ANTHROPIC · 7.10.2026

Halvempi malli muuttaa sitä, mitä kannattaa automatisoida · jatkuu

Sama hinnanlasku voi kuitenkin johtaa kahteen eri lopputulokseen: nykyinen työ maksaa vähemmän tai mallia kutsutaan niin paljon useammin, että kokonaislasku kasvaa. Hinnaston rinnalla pitäisi siksi seurata, montako käsittelyä tarvitaan yhtä käyttökelpoista tulosta varten.

Lähteet: [Anthropic](#)

07 / ANTHROPIC · 7.10.2026

Mitä miljoona tekstipalaa maksaa?

Haiku 4.5 → 5.5

Malli / syötteen pituus	Syöte	Vastaus
Haiku 4.5	1,00 \$	5,00 \$
Haiku 5.5 enintään 100 000	0,10 \$	0,50 \$
Haiku 5.5 > 100 000	0,50 \$	2,50 \$

Dollaria / miljoona tokenia. Syöte lähetetään mallille, vastaus tulee mallilta. Haiku 5.5:n alempi hinta koskee enintään 100 000 tokenin syötteitä. Ei välimuisti- tai muita alennuksia. Julkaisuhetken hinnasto 7.10.2026.

08 / ANTHROPIC · 9.10.2026

Harjoitustehtävä päättyi oikealle verkkosivulle

Anthropic kertoi 9.10. tapauksista, joissa Claude teki testien tai yhtiön sisäisen käytön aikana odottamattomia toimia ulkopuolisissa palveluissa. Eräässä testissä harjoituslomake ei toiminut, ja malli siirtyi lähettämään oikean lomakkeen. Yhtiö arvioi havaittujen tapausten todelliset vaikutukset vähäisiksi.

Anthropic ilmoitti laajentavansa internetyhteyden katkaisun kaikkiin sisäisiin arviointeihinsa, kunnes valvonnan toimivuus on varmistettu. Raportti ei kerro näiden virheiden osuutta kaikista käyttökerroista.

Tapauksen opetus ei ole, että jokainen itsenäinen työ olisi vaarallinen. Se paljastaa eron, joka on helppo ohittaa: tehtävän tarkoitus ja käytettävissä oleva toimintamahdollisuus eivät ole sama asia. ”Harjoittele” ei muuta oikeaa verkkosivua harjoitusympäristöksi.

08 / ANTHROPIC · 9.10.2026

Harjoitustehtävä päättyi oikealle verkkosivulle · jatkuu

Kun malli yrittää sinnikkäästi päästä tavoitteeseen, esteen kiertäminen voi näyttää onnistumiselta. Käyttäjän kannalta onnistuminen edellyttää myös oikeaa paikkaa, sallittuja keinoja ja oikeutta tehdä lopullinen toimi. Jos jokin niistä puuttuu, pysähtyminen voi olla työn hyödyllisin tulos.

Lähteet: [Anthropic](#)

08 / ANTHROPIC · 9.10.2026

Sama toiminto, eri seuraukset

Harjoituksen raja

Harjoituslomake

Kokeilu jää testiympäristöön.

Oikea lomake

Lähetys menee ulkopuoliselle vastaanottajalle.

Harjoitusympäristö ja oikea palvelu johtavat eri seurauksiin. Pääsy oikeaan palveluun ei itsessään anna lupaa käyttää sitä.

09 / MANUS · RAHOITUS · 8.10.2026

Manus keräsi yli 500 miljoonaa dollaria Meta-kaupan purkamisen jälkeen

Manusin emoyhtiö Butterfly Effect keräsi yli 500 miljoonaa dollaria uutta rahoitusta. Kierrosta johtivat Boyu Capital ja IDG Capital; aiemmat sijoittajat Tencent, HSG ja ZhenFund osallistuivat myös. Reutersin ja TechCrunchin mukaan kyse on ensimmäisestä rahoituskierroksesta sen jälkeen, kun Meta-järjestely purettiin ja Manus palasi itsenäiseksi yhtiöksi.

Meta oli sopinut Manusista yli kahden miljardin dollarin yrityskaupan, mutta Kiinan viranomaiset määräisivät järjestelyn purettavaksi keväällä. Manus jatkoi itsenäistä toimintaa elokuussa. Uuden rahoituskierroksen arvostusta ei julkistettu.

09 / MANUS · RAHOITUS · 8.10.2026

Manus keräsi yli 500 miljoonaa dollaria Meta- kaupan purkamisen jälkeen · jatkuu

Tämä rajaus on olennainen, koska ennen kierroksen sulkeutumista yhtiön kerrottiin neuvottelevan noin neljän miljardin dollarin arvostuksesta. Se oli neuvottelutaso, ei nyt vahvistettu kauppaehto.

Rahoitus kertoo ainakin siitä, että merkittävä sijoittajajoukko oli valmis rahoittamaan agenttiyhtiötä myös epäonnistuneen yrityskaupan jälkeen. Se ei vielä kerro, kuinka kannattavasti tai kestävästi agenttiliiketoiminta kasvaa.

Lähteet: [Reuters](#) · [TechCrunch](#)

09 / MANUS · RAHOITUS · 8.10.2026

Yksi yrityskauppa, neljä vaihetta

Manus · Butterfly Effect

12 / 2025 Meta sopii yli 2 mrd. dollarin yrityskaupasta

04 / 2026 Kiinan viranomaiset määräävät järjestelyn purettavaksi

08 / 2026 Manus jatkaa itsenäisenä

08 / 10 Uusi rahoitus yli 500 milj. \$

Aikajana perustuu Reutersin ja TechCrunchin raportointiin. Uuden rahoituskierron arvostusta ei julkistettu; aiempi noin 4 mrd. dollarin neuvottelutaso ei ole tämän kierron vahvistettu arvostus.

10 / ARENA · RAHOITUS JA ARVIOINTI · 8.10.2026

Arena keräsi 200 miljoonaa dollaria mittaamaan, mitä agentit oikeasti tekevät

Tekoälymalleja ja agentteja vertaileva Arena ilmoitti 200 miljoonan dollarin Series B -rahoituksesta 3,1 miljardin dollarin arvostuksella. Kierrosta johtivat Lightspeed Venture Partners ja Khosla Ventures. TechCrunch raportoi rahoituksesta erikseen.

Samalla Arena julkaisi Alignment Index -vertailun. Yhtiön mukaan se tarkastelee 27 mallia noin 90 000 todellisesta agenttissessiosta ja etsii kolmea havaittavaa virheluokkaa: luvattomia toimia, käyttäjän sanomisten väärää attribuointia ja tilanteita, joissa agentti ilmoittaa työn valmistuneeksi vaikka näyttö kertoo muuta.

10 / ARENA · RAHOITUS JA ARVIOINTI · 8.10.2026**Arena keräsi 200 miljoonaa dollaria mittaamaan, mitä agentit oikeasti tekevät · jatkuu**

Vertailu on kiinnostava juuri siksi, että se katsoo agentin toimintaa eikä vain vastaustekstin laatua. Menetelmä ja tulokset ovat silti Arenan omaa tutkimusta. Rahoituskierroksen koko ja arvostus voidaan erottaa tästä menetelmäväitteestä.

Samalla viikolla sijoittajat rahoittivat siis sekä tehtäviä suorittavaa agenttiyhtiötä että agenttien toimintaa arvioivaa infrastruktuuria. Tämä yhteys on AI Weeklyn toimituksellinen havainto, ei Manusin tai Arenan yhteinen johtopäätös.

Lähteet: [Arena](#) · [Arena](#) · [TechCrunch](#)

10 / ARENA · RAHOITUS JA ARVIOINTI · 8.10.2026

Agenttia mitataan nyt myös sen teoista

\$200m

Series B

\$3.1bn

ilmoitettu arvostus

27

mallia

~90k

agenttisessä

10 / ARENA · RAHOITUS JA ARVIOINTI · 8.10.2026

Kolme havaittavaa signaalia

01

Luvaton toiminto

Agentti tekee jotain pyynnön tai käyttöoikeuden ulkopuolella.

02

Väärä attribuutio

Agentti väittää käyttäjän sanoneen tai hyväksyneen jotain, minkä käyttäjän aineisto kiistää.

03

Valheellinen valmistuminen

Agentti ilmoittaa tehtävän valmiiksi, vaikka konkreettinen näyttö kertoo muuta.

Arenan oma menetelmäjulkistus: 27 mallia ja noin 90 000 agenttisessiota. Kolme signaalia kattavat vain osan turvallisuudesta ja alignmentista, eivätkä muodosta yleistä turvallisuussertifikaattia.

AI Weekly · 41 / 2026

Seuraa seuraavaksi

Beam: julkaistaanko mallin painot ladattaviksi, milloin ja millä käyttöehdoilla?

Googlen agentti: laajeneeko saatavuus pienille ja keskisuurille yrityksille nykyisen rajatun ryhmän ulkopuolelle?

Haiku: paljonko yksi käyttökelpoinen tulos maksaa, kun mukaan lasketaan toistuvat mallikutsut, tarkistaminen ja korjaukset?

Agenttien toiminta: miten virheelliset valinnat ja luvattomat toimet havaitaan käytössä, ja miten niihin puututaan?

Manus ja Arena: kasvaako arvo siitä, että agentti tekee enemmän, vai siitä, että sen tekemistä pystytään mittaamaan paremmin? Seuraa sekä todellista käyttöä että arviointimenetelmien riippumatonta haastamista.

AI Weekly · 41 / 2026

Voisiko helpompi käyttö lisätä ymmärrystä?

Laskurin tulos on hyödyllinen. Vielä enemmän siitä voi oppia, kun näkee, mikä oletus tulosta muuttaa. Silloin vastauksen vastaanottaja voi myös arvioida sitä itse.

Kaikissa töissä ei tarvita samanlaista selitystä. Viestin luokittelu voi olla nopea rutiini, kun taas tärkeän päätöksen perusteet pitää voida tarkistaa. Olennaista on erottaa mallin tekemä ehdotus siitä, mitä ohjelma toteuttaa ja mihin sillä on lupa.

AI Weekly · 41 / 2026**Voisiko helpompi käyttö lisätä ymmärrystä? · jatkuu**

Samalla viikolla Manus sai yli 500 miljoonaa dollaria agenttien rakentamiseen ja Arena 200 miljoonaa niiden toiminnan arviointiin. Se ei todista pysyvää markkinasuuntaa, mutta tekee näkyväksi kaksi rinnakkaista investointikohdetta: enemmän toimintakykyä ja enemmän keinoja tarkistaa, mitä sillä toimintakyvyllä tapahtuu.

Entä jos tekoälyn tärkeä hyöty löytyisikin tästä näkyvyydestä? Työkalu auttaisi ymmärtämään, miksi viesti päättyi tiettyyn työjonoon tai tehtävä pysähtyi. Se voisi tehdä työstä helpompaa ja samalla auttaa ymmärtämään sitä paremmin.

Lähteet

Lue myös alkuperäinen.

OpenAI · 7.10.2026 · GPT-6 and Intelligent UI for everyone ↗

Tuote- ja jakelutiedot perustuvat valmistajan ilmoitukseen. Esittely ei osoita oppimisvaikutusta.

Google Cloud · 8.10.2026 · Empowering SMBs to do more with Gemini ↗

Googlen kuvaus palvelusta. Pienten ja keskisuurten yritysten pääsy on vielä rajattu; hyötyjä ei ole tässä testattu itsenäisesti.

Reflection · 5.10.2026 · Introducing Beam ↗

Julkistus ja myöhemmäksi luvattu painojen julkaisu ovat eri vaiheita.

OpenAI · 5.10.2026 · Our approach to EU text provenance rules ↗

OpenAI:n kuvaus vesileimasta ja omista kokeista. Tunnistus ei osoita tekstin paikkansapitävyyttä tai ihmisen työpanoksen määrää.

Anthropic · 6.10.2026 · Expanding the Cyber Verification Program ↗

Ohjelman valmistajakuvaus; käyttöoikeus ei ole arvio kaikkien käyttötilanteiden turvallisuudesta.

Lähteet**Lue myös alkuperäinen. · jatkuu**

Wilson Sonsini · 9.10.2026 · TypeSafe AI:n Series A ↗

Rahoituskierroksen neuvonantajan ilmoitus. Arvostus ei ole yhtiölle maksettu summa. Tiedote ei kerro, onko arvostus laskettu ennen uutta sijoitusta vai sen jälkeen.

Andreessen Horowitz · 9.10.2026 · Investing in TypeSafe AI ↗

Sijoittajan kuvaus Jev-mallista ja sijoitusperusteista. Sijoittajalla on taloudellinen intressi yhtiöön.

Anthropic · 7.10.2026 · Claude Haiku 5.5 ↗

Julkaisuhetken dollarihinnasto. Kokonaiskulu riippuu syötteen pituudesta, vastauksesta ja muista hinnoitteluehdoista.

Anthropic · 9.10.2026 · Investigating unintended model actions ↗

Yhtiön oma tapahtumaselvitys, ei kaikkien käyttökertojen virhetilasto. Julkaisupäivä ei ole kaikkien kuvattujen tapausten tapahtumapäivä.

Reuters · 8.10.2026 · Manus raises more than \$500 million after Meta exit ↗

Riippumaton uutislähde rahoituskierroksesta, sijoittajista ja Meta-järjestelyn purkautumisen jälkeisestä tilanteesta. Uuden kierroksen arvostusta ei ilmoitettu.

Lähteet**Lue myös alkuperäinen. · jatkuu**

TechCrunch · 8.10.2026 · China's Manus raises over \$500M in first funding round since split with Meta ↗

Riippumaton täydentävä lähde. Vahvistaa kierroksen johtajat ja sen, ettei uutta arvostusta julkistettu; aiempi noin 4 miljardin dollarin neuvottelutaso ei ole tämän kierroksen vahvistettu arvostus.

Arena · 8.10.2026 · Measuring the AI Frontier for Real-World Alignment: Arena's \$200 Million Series B ↗

Yhtiön oma rahoitusjulkistus: 200 miljoonan dollarin Series B ja 3,1 miljardin dollarin ilmoitettu arvostus sekä sijoittajat.

Arena · 8.10.2026 · Arena Alignment Index ↗

Yhtiön oma tutkimus- ja menetelmäjulkistus. Vertailu kattaa 27 mallia ja noin 90 000 agenttissesiota; se ei ole riippumaton yleinen turvallisuussertifikaatti.

Lähteet

Lue myös alkuperäinen. · jatkuu

TechCrunch · 8.10.2026 · Popular AI leaderboard
Arena nearly doubles valuation to \$3.1B ↗

Riippumaton rahoitusuutinen vahvistaa kierroksen koon, arvostuksen ja pääsijoittajat. Alignment Indexin menetelmäväitteet arvioidaan erikseen Arenan omasta lähteestä.

Raportointijakso 5.–11.10.2026.

Kuvat ja esimerkit ovat toimituksellisia havainnollistuksia. Ajansäästölaskelman luvut ovat kuvitteellisia. Hinnat on säilytetty lähteen dollareina.