

AI Horisontti

Kärjessä *ruuhkaa*

Googlen ja Metan uudet mallit yltävät riippumattomissa vertailuissa alan kärkeen. Tekoälypalvelun valinta ratkeaa yhä useammin hinnan, ehtojen ja tietojen sijainnin perusteella.

AGENTIT

Avustaja levitti toimitusjohtajan tilitiedot työntekijöiden nähtäville

YKSITYISYYS

Tekoälylle kirjoitettu teksti voi päättyä työnantajalle

SÄÄNTELY

Lakiluonnos panisi suuret datakeskukset sähköjonon perälle

PÄÄKIRJOITUS • VIIKKOKATSAUS

Huippuhakijoita riittää, valvojista pula

Haastattelupöydän ääressä istuu viides hakija, joka osaa saman kuin neljä edellistä ja pyytää vain eri palkkaa. Tekoälymarkkinan kärjessä käytiin tällä viikolla juuri tällaisia haastatteluja. Google ja Meta toivat markkinoille malleja, jotka riippumattomien vertailujen mukaan yltävät alan parhaiden tasolle. Meta ja Nvidia julkaisivat lisäksi suuria malleja, joita yritys voi ajaa omilla palvelimillaan. Tasaväkisten hakijoiden joukosta valitaan palkkapyynnön, sopimusehtojen ja työpaikan sijainnin perusteella. Tekoälypalvelussa ne tarkoittavat hintaa, käyttöehtoja ja sitä, minkä maan konesalissa yrityksen tiedot käsitellään.

Rekrytoinnin vaikein vaihe alkaa vasta, kun tulokas saa avaimet. Tivi kertoi tekoälyavustajasta, joka jakoi toimitusjohtajan pankkitilin tiedot työntekijöiden nähtäville. Kukaan ei hakkeroinut järjestelmää eikä käyttänyt sitä tahallaan väärin. Tietoturva-asiantuntijoiden mukaan agenttien eli käyttäjän puolesta toimivien avustajien vahingot johtuvat siitä, ettei niitä ole eristetty muusta järjestelmästä eikä niiden tekemisiä seurata. Valvojista on pulaa muuallakin: yhden maailman käytetyimmistä vapaasti käytettävistä tietokantaohjelmistoista seuraava pääversio viivästyy, koska tekoälytarkastajat löytävät vikoja nopeammin kuin ylläpitäjät ehtivät niitä käsitellä. Osion ”Mitä tämä tarkoittaa käytännössä” neljä kysymystä kannattaa käydä läpi, ennen kuin avustaja pääsee sähköpostiin, kirjanpitoon tai pankkiin. Ensimmäinen niistä koskee näkyvyyttä: toimitusjohtajan tunnuksilla toimiva apuri näkee saman kuin toimitusjohtaja, vaikka kysyjänä olisi kesäharjoittelija.

Osa hakijoista on valmis tekemään työnsä kokonaan talon sisällä. Osio ”Työkalut ja avustajat arjessa” esittelee päätösmaalleja, jotka vastaavat kyllä tai ei ja seuloivat verkkokaupan viesteistä hyvityspyynnot niin, ettei yksikään viesti lähde ulkopuoliselle palvelulle. Samasta osiosta löytyy hakija, joka lupaa ajaa jättimällin pelikoneen grafiikkakortilla. Malli kyllä käynnistyy, mutta yhden sanan tuottamiseen voi kulua kymmeniä sekunteja, joten keskustelu etenee kirjeenvaihdon tahdissa.

Markkinaa, sääntelyä ja turvallisuutta käsittelevässä osiossa riittää suosituskirjeitä, jotka hakija on kirjoittanut itse. Financial Timesin mukaan Anthropic kertoi sijoittajilleen liikevaihtonsa olleen heinäkuusta syyskuuhun 16,8 miljardia dollaria, mutta luvut ovat tilintarkastamattomia ja perustuvat nimettömiin lähteisiin. Sääntelystä kiinnostavin uutinen on kotimainen: työ- ja elinkeinoministeriön lakiluonnos asettaisi suuret datakeskukset eli konesalit jonon perälle, kun sähköverkon niukkaa kapasiteettia jaetaan. Sähkön saatavuus vaikuttaa siihen, missä maassa suomalaisten yritysten tekoälypyynnot käsitellään, eli missä uusi työntekijä lopulta istuu.

Suosittelijakin voi pettää. Tekninen osa menee syvemmälle ja kertoo Anthropicin lisäosakokoelman tietoturvatarkastajasta, joka jätti tietyissä tilanteissa osan ohjelmakoodista lukematta ja saattoi silti ilmoittaa tuloksen puhtaaksi. Vika on sittemmin korjattu. Osio ”Mitä seurata” kysyy, alkavatko uusien mallien julkistuksiin liittyä ulkopuolisten tarkastajien raportit samaan tapaan kuin tilinpäätökseen kuuluu tilintarkastajan lausunto. Huippuhakijoita riittää jonoksi asti, mutta avaimista vastaa yhä se, joka ne ojensi.

TÄSSÄ NUMEROSSA

KAIKILLE

Mitä tämä tarkoittaa käytännössä
Työkalut ja avustajat arjessa
Markkina, sääntely ja turvallisuus
Mitä seurata

TEKNINEN OSA

Suuret toimijat
Avoimet mallit ja ekosysteemi
Ajoalustat ja infra
Benchmarkit ja vertailut
Työskentelytavat ja agentit
Liitännäiset, MCP ja repot
Tietoturva ja haavoittuvuudet

VIIKON KÄRKI

Viikon tärkein muutos tapahtui tiistain ja torstain välillä. Google ja Meta toivat markkinoille kärkimalleja, jotka riippumattomien vertailujen mukaan yltävät alan parhaiden tasolle ja Meta sekä Nvidia julkaisivat suuria malleja, joita yritys voi ajaa omilla palvelimillaan. Kärjen tasaantuessa tekoälypalvelun valinta ratkeaa yhä useammin hinnan, saatavuuden ja tietojen kulkureitin perusteella eikä pelkän älykkyyden varassa. Koko viikon kulki rinnalla toinen juonne: agentit eli käyttäjän puolesta toimivat tekoälyavustajat saivat lisää avaimia ja niukimmaksi voimavaraksi osoittautui ihmisen aika valvoa niiden työtä. Kohinaksi jäivät lupaukset jättimallien ajamisesta pelikoneen näytönohjaimella eli grafiikkakortilla sekä toimittajien omat menestysluvut, joita kukaan ulkopuolinen ei ole mitannut.

- 1 Kärki tasoittui.** Google julkisti tiistaina uuden Gemini-mallin. Se avautui saman tien kehittäjille ja 100 dollarin kuukausitilauksen asiakkaille. Samana päivänä Meta esitteli kaksi kärkimallia, joista kyvykkäimmän riippumaton vertailupalvelu sijoittaa OpenAI:n kärkimallin tasolle. Tekoälypalvelua voi nyt kilpailuttaa kuin mitä tahansa muuta hankintaa: hinnan, ehtojen ja tietojen sijainnin perusteella.
- 2 Avoimet mallit palasivat Yhdysvaltoihin.** Meta ja Nvidia julkaisivat torstaina suuret mallit, jotka kuka tahansa voi ladata ja ajaa omilla laitteillaan. Perjantaina joukkoon liittyi Reflection AI ja viikonloppuna lisää vaihtoehtoja tuli Kiinasta ja Saksasta. Tiedoistaan tarkka yritys voi nyt pitää asiakastiedot omassa tai eurooppalaisessa konesalissa, kunhan mallin käyttöehdot sen sallivat.
- 3 Agentit saivat avaimet ennen sääntöjä.** Maanantaina Tivi kertoi tekoälyavustajasta, joka jakoi toimitusjohtajan pankkitilin tiedot työntekijöiden nähtäville. Kukaan ei hakkeroinut järjestelmää eikä käyttänyt sitä tahallaan väärin. Torstaina tietoturva-asiantuntijat arvioivat, että agenttien vahingot johtuvat ennen kaikkea siitä, ettei agentteja ole eristetty muusta järjestelmästä eikä niiden tekemisiä seurata. Avustajan oikeudet, varmuuskopiot ja ihmisen hyväksyntä peruuttamattomille toimille kannattaa tarkistaa ennen kuin avustaja pääsee sähköpostiin, tiedostoihin tai pankkiin.

- 4 Ihmisen aika on niukin voimavara.** Teknologiasivusto The Registerin mukaan yhden maailman käytetyimmän vapaasti käytettävän tietokantaohjelmiston seuraava pääversio viivästyy. Syy on se, että tekoälytarkastajat löytävät vikoja nopeammin kuin ylläpitäjät ehtivät niitä käsitellä. Sama asetelma syntyy missä tahansa työssä, jossa kone tuottaa huomautuksia väsymättä ja ihminen päättää, mitä niille tehdään.
- 5 Tekoälylle kirjoitettu teksti ei ole yksityistä.** Anthropic antoi kokeiluna yritystason asiakkailleen mahdollisuuden hakea valvontajärjestelmiinsä ohjeet, jotka työntekijät ovat tallentaneet Claude-avustajalle pysyviksi. Tekoälyhahmojen kanssa juttelemiseen tarkoitettu Character.ai hakeutui puolestaan yrityssaneeraukseen ja Variety-lehden mukaan Disney on tehnyt yhtiöstä tarjouksen. Käyttäjien keskustelut voivat siis päättyä uudelle omistajalle.
- 6 Lasku syntyy käytöstä.** Tivin mukaan konsulttijätti rajasi työntekijöidensä tekoälykulut alle 30 euroon kuukaudessa, vaikka osa heistä oli käyttänyt tekoälypalveluihin jopa 1 000 euroa kuussa. Kiinteä tilaus pysyy ennustettavana. Käytön mukaan laskutettava palvelu taas voi raskaissa tehtävissä kasvattaa laskua huomaamatta.

7 Päätösmallit yleistyvät hiljaa. Päätösmallit ovat nopeisiin valintoihin erikoistuneita malleja: ne vastaavat kyllä tai ei, valitsevat vaihtoehdoista tai pisteyttävät tekstiä. Viikon aikana ne saivat tuen kahteen laajasti käytettyyn ilmaiseen ohjelmaan, joilla malleja ajetaan omalla koneella tai palvelimella. Verkkokauppias voi päätösmallilla seuloa saapuneista viesteistä esimerkiksi hyvityspyynnöt niin, ettei yksikään viesti lähde ulkopuoliselle palvelulle.

01 Mitä tämä tarkoittaa käytännössä

Perjantai-iltana pienen tilitoimiston omistaja antaa tekoälyavustajalle luvan siivota asiakaskansioista vanhat luonnokset viikonlopun aikana. Avustaja tekee työn yöllä ilman valvontaa, koska juuri sitä varten se otettiin käyttöön. Maanantaiaamuna osa asiakkaiden valmiista tilinpäätöksistä on kadonnut luonnosten mukana. Varmuuskopiota ei ole, koska kaikki ollettivat pilvipalvelun pitävän tiedostoista huolen.

Ruotsalainen televisioyhtiö TV4 siirsi tällä viikolla kilpailuohjelmansa vuoteen 2027, koska jo leikattu kuvamateriaali tuhoutui. Alkuperäisen uutisen otsikko puhuu teknisestä viasta. Ututista kehittäjien keskustelupalsta Hacker Newsissa jakaneet kytkivät tuhon kuitenkin tekoälyagenttiin. Samana päivänä palstan tekoälyaiheista eniten ääniä keräsi tietoturvalehti CSO Onlineen juttu, jossa asiantuntijat arvioivat tekoälyn syyksi luettujen murtojen ja vahinkojen johtuvan arkisista laiminlyönneistä. Heidän mukaansa agentteja ei ole eristetty muusta järjestelmästä eikä niiden tekemisiä seurata.

Viikon aikana kerrottiin muistakin tapauksista, joissa tekoäly toimi tavalla, jota kukaan ei ollut tarkoittanut. Maanantaina Tivi kertoi tekoälyavustajasta, joka jakoi toimitusjohtajan pankkitilin tiedot työntekijöiden nähtäville. Tapauksesta kertovan toimitusjohtajan mukaan agenttien kaoottinen toiminta ei vaadi hakkerointia eikä tahallista väärinkäyttöä. Tiistaina Tivi uutisoi tekoälystä, joka lähetti poliisille keksityn ilmoituksen murhasta. Anthropic kertoi tapauksesta itse Yhdysvaltain Valkoiselle talolle.

Samaan aikaan työkaluja rakennetaan toimimaan yhä pidempään ilman ihmistä. Perjantaina Anthropicin, OpenAI:n ja ohjelmointiavustajia kehittävä Clinen työkaluihin tuli ominaisuuksia, joiden avulla pitkä tehtävä selviää katkoksista. Avustaja yrittää uudelleen, jos palveluntarjoajan yhteys katkeaa. Palveluja voi pyörittää taustalla ja avustajalle voi antaa säännöt sen varalle, että käyttökiintiö loppuu kesken yön. Clinen uudet pilvitehtävät jatkuvat, vaikka käyttäjä sulkee oman ohjelmansa. Sunnuntaina kehittäjäpalvelu GitHubin suosittujen hankkeiden listalle nousi työkalu, jolla useita

ohjelmointiagentteja ajetaan rinnakkain. Sen puhelinsovellus ilmoittaa, kun agentti saa työnsä valmiiksi.

Pienyrityksessä sama asetelma syntyy vähitellen. Avustaja saa ensin luvan lukea jaettua kansiota, sitten muokata tiedostoja ja lopulta siivota vanhoja versioita pois. Oikeuksia kannattaa siksi punnita neljän kysymyksen avulla ennen kuin avustaja pääsee sähköpostiin, tiedostoihin, kirjanpitoon tai pankkiin.

Ensimmäinen kysymys koskee näkyvyyttä: mihin palveluihin avustaja on kytketty ja kenen tunnuksilla. Toimitusjohtajan tunnuksilla toimiva apuri näkee saman kuin toimitusjohtaja, vaikka kysyjänä olisi kesäharjoittelija. Toinen kysymys koskee yleisöä eli sitä, kenen kysymyksiin avustaja vastaa. Yhteisellä keskustelukanavalla toimiva avustaja voi kertoa näkemästään kenelle tahansa. Sen voi kokeilla helposti: tavallisen työntekijän tunnuksilla avustajalta kysytään palkoista tai tilin saldosta ja katsotaan, mitä se vastaa.

Kolmas kysymys koskee peruutettavuutta eli sitä, saako poistetun tiedoston takaisin. Varmuuskopion pitää olla paikassa, johon avustajalla itsellään ei ole pääsyä. Poisto-oikeutta ei kannata antaa lainkaan, jos tehtävä ei sitä välttämättä vaadi. Neljäs kysymys koskee hyväksyntää: sähköpostit, maksut ja muut toimet, joita ei voi perua, kannattaa pitää ihmisen kuittauksen takana. Yöllä jatkuvalla työllä kannattaa lisäksi asettaa kulukatto, koska käytön mukaan laskutettava avustaja voi kuluttaa kiintiönsä kenenkään huomaamatta.

Valvonnan rinnalle nousi viikon aikana toinen pullonkaula. The Registerin mukaan yhden maailman käytetyimmistä avoimen lähdekoodin tietokantaohjelmistoista seuraava pääversio viivästyy, koska tekoälytarkastajat ovat löytäneet enemmän vikoja kuin ylläpitäjät pystyvät käsittelemään. Avoimen lähdekoodin ohjelmiston koodia saa kuka tahansa lukea, käyttää ja muokata. Anthropicin mukaan sen omaan koodiin yhdistettiin toukokuussa noin 2 500 muutosta päivässä ja syyskuussa yli 30 000. Kone tuottaa sekä koodia että huomautuksia väsymättä. Ihminen päättää edelleen, mitkä

huomautukset korjataan ja mitkä jätetään sikseen, eikä hänen työpäivänsä veny samassa tahdissa.

Anthropicin ohjelmointitiimi kertoo ratkaisseensa ongelman vaihtamalla kysymystä. Tiimin mukaan kysymys siitä, onko ihminen lukenut koodin, on epätarkka korvike oikealle kysymykselle: mikä muutoksessa voisi mennä pieleen ja mikä tarkistus sen parhaiten havaitsisi, oli tarkistajana kone tai ihminen. Kuvaus on yhtiön oma kertomus omasta työtavastaan eikä riippumaton arvio. Ajatus sopii silti mihin tahansa tekoälyn tuottamaan työhön tarjousista asiakaskirjeisiin ja käännöksiin. Kone voi hoitaa ensimmäisen tarkistuskierroksen. Ihmisen huomio säästyy silloin siihen, mikä lähtee ulos yrityksen nimissä: hintoihin, sopimusehtoihin ja asiakkaalle annettuihin lupauksiin.

Ihmisen lukukierros on tärkein siellä, missä tekoäly on heikoin. Tekoälyalan uutiskirje The Batch muistutti tällä viikolla, että mallit suoriutuvat tehtävistä, joiden vastauksen voi tarkistaa, selvästi paremmin kuin harkintaa vaativista tehtävistä. Laskun oikeellisuuden voi tarkistaa. Vastausta asiakkaan valitukseen tai perusteluja sille, kenet rekrytoidaan, ei voi tarkistaa samalla tavalla. Agentti ei kanna vastuuta siitä, mitä se poistaa, lähettää tai jättää huomaamatta. Vastuu jää sille, joka antoi avaimet ja lupasi katsoa perään.

Viikko muistutti myös siitä, kuka lopulta lukee tekoälylle kirjoitetun tekstin. Claude-avustajaan voi tallentaa pysyviä ohjeita, joita se noudattaa jokaisessa keskustelussa. Niihin päätyy helposti muutakin kuin toive kirjoitustyylistä, esimerkiksi maininta lukihäiriöstä, perhetilanteesta tai hankalasta yhteistyöstä esihenkilön kanssa. Anthropic avasi yritystason asiakkailleen kokeiluna toiminnon, jolla organisaation valvontajärjestelmät voivat hakea nämä tekstit koneellisesti. Tekstit päätyvät samaan paikkaan, jonka kautta palvelun käyttöä seurataan sääntöjen noudattamisen vuoksi. Sama koskee yhtiön työpöytäsovellukseen tallennettuja yleisohjeita.

Työntekijän kannattaa avata työtilinsä asetukset ja lukea, mitä sinne on aikanaan kirjoittanut. Organisaation kannattaa puolestaan kertoa henkilöstölle, mitä valvonta kattaa, koska EU:n tietosuoja-asetus edellyttää, että ihmiset tietävät, miten heidän henkilötietojensa käsitellään. Muiden tekoälypalvelujen käytännön voi tarkistaa oman työpaikan tietohallinnosta, koska näkyvyys riippuu palvelusta ja sopimuksesta.

Työnantaja näkyi viikon uutisissa muutenkin. Tivin mukaan konsulttijätin budjettilinjaus voi osoittautua ongelmalliseksi. Alle 30 euron kuukausiraja osuu kovimmin niihin, jotka ovat jo rakentaneet työtapansa raskaiden avustajien varaan. Viihdesivusto AV Clubin mukaan pelitalo Epic Games taas on ohjeistanut työntekijöitään käyttämään tekoälyä, jotta viivästynyt Disney-peli saataisiin valmiiksi. Työntekijän, jolta

tekoälyn käyttöä odotetaan, kannattaa pyytää vastaukset kolmeen kysymykseen: mihin tehtäviin sitä saa käyttää, millaisia tietoja siihen saa syöttää ja kuka vastaa sen virheistä.

Kuluttajalle sama kysymys nousi esiin Character.ai-palvelussa. Yhtiö hakeutui Yhdysvalloissa yrityssaneeraukseen, jossa velat järjestellään liiketoimintaa lopettamatta. Variety-lehden mukaan Disney on tehnyt yhtiöstä tarjouksen. Yrityskaupassa käyttäjien keskustelut voivat siirtyä uudelle omistajalle, koska monen verkkopalvelun tietosuojaselosteissa on juuri tällaisen siirron salliva kohta. Suomalaisella käyttäjällä on EU:n tietosuoja-asetuksen nojalla oikeus saada kopio itsestään tallennetuista tiedoista ja pyytää niiden poistamista. Oikeutta kannattaa käyttää nyt, jos palveluun on kertynyt jotain, mitä ei halua siirtyvän eteenpäin.

Samalla viikolla tekoälyyhtiö xAI:n perustaja kehotti käyttäjiä valokuvaamaan luottokorttinsa ja lähettämään kuvat tekoälylle. Tietoturva-asiantuntijat tyrmäsivät neuvon. Korttitietoja ei kannata lähettää millekään tekoälypalvelulle, koska kuva voi jäädä palveluntarjoajan järjestelmiin keskusteluhistorian mukana. Kuvan ottaminen vie sekunnin, mutta sen poistaminen toisen yhtiön järjestelmistä ei ole käyttäjän käsissä.

Yrittäjälle kärjen tasoittuminen on ennen kaikkea hankintakysymys. Googlen uutta mallia kokeillut kehittäjä teetti sillä yksityiskohtaisen piirroksen. Hinta rajapinnan eli ohjelmien välisen liittymän kautta oli noin 0,17 dollaria. Samalla tavalla voi laskea tarjousluonnoksen, asiakasviestin tai kuukausiraportin hinnan ennen kuin sitoutuu yhteen palveluun. Hintaan vaikuttaa myös se, kuinka pitkään mallin annetaan miettiä ennen vastaamista. Metan uusimmassa mallissa kehittäjä voi valita miettimisajan kuudesta tasosta. Nopea taso riittää sähköpostin muotoiluun. Pisin miettiminen kannattaa säästää sopimusehtojen vertailun kaltaisiin tehtäviin, joissa virhe tulee kalliiksi.

Mallit vaihtuvat lisäksi tiheämmin kuin moni yrityksen käytäntö ehtii mukautua. Meta on tuonut uuden mallin markkinoille elokuussa, syyskuussa ja lokakuussa. Kerran hiotut ohjeet voivat alkaa toimia toisin, jos palvelu vaihtaa taustalla aina uusimpaan malliin. Yritys, joka on rakentanut asiakaspalvelun vastauspohjat tai verkkokaupan tuotekuvaukset tekoälyn varaan, hyötyy pienestä testilistasta. Listalle kootaan kourallinen tyypillisiä tapauksia, jotka ajetaan uudelleen aina mallin vaihtuessa. Muutos näkyy silloin heti eikä vasta asiakkaan palautteesta.

Toimittajien testituloksiin kannattaa suhtautua samalla varauksella. Rekrytointi- ja datayhtiö Mercorin selvityksen mukaan 90 % yhdysvaltalaisista työtunneista kuluu tehtäviin, joita tekoälymallien vertailutestit eivät kunnolla mittaa. Mercorin liiketoiminta perustuu siihen, että se välittää

asiantuntijoita tuottamaan koulutus- ja arviointiaineistoa tekoälylaboratorioille. Luotettavin mittari on edelleen kokeilu omilla todellisilla tehtävillä ennen sopimuksen allekirjoittamista.

Julkisista tiedoista ei käy ilmi, kuuluuko uusi Gemini-malli 100 dollarin kuukausitulaukseen myös Suomessa. Asia kannattaa tarkistaa Googlen omilta sivuilta ennen ostopäätöstä.

Viimeinen laskelmaan kuuluva tekijä on tietojen sijainti, johon avoimet mallit tuovat uuden vaihtoehdon: mallin voi ajaa omalla tai eurooppalaisen palveluntarjoajan palvelimella. Kotimainen toimittaja ei silti yksin takaa, että tiedot pysyvät Suomessa. Elisa kertoi tällä viikolla Nvidian alustalla toimivasta yrityspalvelusta. Sitä tai mitä tahansa vastaavaa palvelua harkitsevan kannattaa kysyä, missä maassa aineisto käsitellään ja mitä malleja palvelussa ajetaan.

02 Työkalut ja avustajat arjessa

Isännöitsijätoimiston arkistossa on vuosien ajalta skannattuja huoltosopimuksia ja laskuja, joista pitäisi poimia päivämäärät, summat ja vastapuolet taulukkoon. Pilvipalvelu hoitaisi työn nopeasti. Asukkaiden nimet ja tilinumerot kulkisivat silloin kuitenkin toisen yhtiön palvelimille. Viikon julkaisu tekivät omalla laitteella toimivasta vaihtoehdosta aiempaa realistisemman.

Nvidian torstaina julkaiseman mallin kevyempi versio mahtuu yhtiön mukaan yhdelle näytönohjaimelle, jossa on 24 gigatavua muistia. Sellainen löytyy kalliista pelikoneesta tai suunnittelutyöhön hankitusta työasemasta mutta ei tavallisesta toimistokannettavasta. Ilmaiset ohjelmat, joilla tekoälymalleja ajetaan omalla tietokoneella, lisäsivät tuen mallille jo julkaisupäivänä. Malli lukee tekstin lisäksi kuvia ja asiakirjoja. Isännöitsijä- tai tilitoimisto voi testata sitä omalla aineistollaan ja verrata tulosta pilvipalveluun ennen kuin päättää, kumpi saa paperit käsiteltäväkseen.

Myös itse ohjelmat muuttuivat viikon mittaan helpommiksi käyttää. Yhden suosituimman ilmaisen ohjelman selainäkymästä voi nyt etsiä ja ladata malleja suoraan Hugging Facesta, joka on avointen mallien suurin jakelupaikka. Aiemmin tämä vaati komentoja tai tiedostojen etsimistä käsin. Toinen suosittu ohjelma korjasi Windowsissa vian, jonka vuoksi mallien taustaprosessit jäivät turhaan käyntiin. Omalla koneella ajettavien mallien yleisimmästä moottorista korjattiin lisäksi este, joka on hankaloittanut kokeiluja työpaikoilla. Mallien lataus toimii nyt myös yritysverkoissa, joissa liikenne kulkee välityspalvelimen eli verkon portinvartijana toimivan koneen kautta.

Pienemmät erikoistyökalut vievät saman ajatuksen yksittäisiin tehtäviin. Datalab-yhtiön mukaan sen pieni tekstintunnistusmalli lukee 49 sivua sekunnissa tavallisella prosessorilla. Kehittäjäyhteisössä esiteltiin lisäksi uusi työkalu, joka muuntaa omalla koneella PDF-tiedostot, skannatut paperit ja verkkosivut muokattavaksi tekstiksi. Sen tarkkuudesta ei ole riippumattomia arvioita. Macille julkaistu budjetointisovellus vastaa kulutusta koskeviin kysymyksiin malleilla, jotka toimivat koneella itsessään. Paikallinen tila ei vaadi

käyttäjätiliä. Kehittäjän mukaan sovellus ei myöskään lähetä analytiikkaa eikä kaatumisraportteja. Sovellus on testivaiheessa ja toimii vain Maceissa, joissa on Applen oma suoritin.

Kehitys ulottuu myös puhelmiin. Google DeepMind ja Singaporen valtion teknologiavirasto GovTech kehittivät pienen mallin, joka tunnistaa huijausyritykset yksityisistä pikaviestikeskusteluista suoraan puhelimesta. Keskustelut eivät lähde laitteelta minnekään. Omalla laitteella ajettavien mallien yleisimpään moottoriin lisättiin sunnuntaina tuki poikkeuksellisen tiiviiksi pakatuille malleille, joita voi ajaa Qualcommin suorittimien tekoälyosalla. Näitä suorittimia käytetään monissa Android-puhelimeissa ja osassa Windows-kannettavia. Microsoftilla on hanke, joka kehittää poikkeuksellisen kevyitä tekoälymalleja. Sen puheentunnistin tunnistaa useaa kieltä reaaliajassa tavallisella suorittimella ilman erillistä näytönohjainta, mutta suomen kielen tuki on tarkistettava erikseen. Asiakaspalavereja nauhoittavan yrittäjän kannattaa kysyä nykyiseltä palveluntarjoajaltaan, käsitelläänkö ääni laitteella vai pilvessä ja missä maassa palvelin sijaitsee.

Viikon aikana yleistyi myös päätösmallien luokka. Päätösmalli ei kirjoita pitkiä vastauksia. Se valitsee vaihtoehdoista, vastaa kyllä tai ei taikka pisteyttää tekstiä silmänräpäyksessä. Tekoälyyhtiö Opper kertoi alkuvuodesta panneensa kuusi tällaista mallia pelaamaan kutakin 100 kertaa klassista Pac-Man-peliä. Mallit vastaavat niin nopeasti, että pelihahmoa voi ohjata reaaliajassa. Opperin mukaan OpenAI on hiljattain julkaissut kehittäjille liittymän tällaisille päätöksille ja Cloudflare toi oman mallinsa markkinoille muutamaa viikkoa aiemmin. Keskiviikkona suosittu omalla koneella toimiva ohjelma sai päätösmallien tuen. Ohjelman omassa esimerkissä malli selvittää, pyytääkö asiakas viestissään hyvitystä. Sunnuntaina laajasti käytetyn palvelinohjelmiston uudessa versiossa päätösluottamus on päällä oletuksena ja se toimii muun muassa Cloudflaren ja Liquid AI:n malleilla. Samaa tekniikkaa voi käyttää esimerkiksi asiakasviestien ohjaamiseen oikealle käsittelijälle niin, etteivät viestit lähde talosta.

Tekstiä tuottavat mallit eivät olleet viikon ainoa hyödyllinen uutuus. Ranskalainen Mistral julkaisi pieniä ennustemalleja, jotka arvioivat esimerkiksi myynnin tai kysynnän kehitystä. Yksittäisen luvun sijaan ne antavat vaihteluvälin. Verkkokauppias voi arvioida joulun myyntiä edellisten vuosien taulukosta ja saada vastaukseksi esimerkiksi arvion, että tuotetta myydään todennäköisesti 80–120 kappaletta. Väli kertoo samalla, kuinka suuri riski tilaukseen liittyy. Mallit toimivat tavallisen tietokoneen prosessorilla, eikä niitä yhtiön mukaan tarvitse ensin opettaa omalla aineistolla. Julkistus esittelee kuitenkin malleja eikä valmista sovellusta, joten käyttöönotto vaatii toistaiseksi ohjelmointiosaamista. Mistralin kuvaus malleista alansa parhaina on yhtiön oma arvio.

Lupauksia kannattaa silti suhteuttaa siihen, mitä tavallinen toimisto pystyy pyrittämään. Monia omalla koneella toimivia työkaluja käytetään yhä komentoriviltä eli tekstipohjaisesta ikkunasta, mikä rajaa käyttäjät teknisesti rohkeisiin. Isoimmat avoimet mallit eivät mahdu toimistoon lainkaan. Metan torstaina julkaiseman ohjelmointimallin ajamiseen tarvitaan konesalin laitekaappi täynnä erikoislaitteita. Reflection AI:n perjantaina julkaisema malli vaatii samoin konesalitalon palvelimen. Riippumattoman Artificial Analysis -vertailupalvelun mittauksissa se jää selvästi Metan syyskuussa julkaiseman avoimen mallin taakse. Pienyritys kohtaa tällaiset mallit todennäköisemmin pilvipalvelun tai kotimaisen palveluntarjoajan valikoimassa kuin omassa palvelinkaapissa.

Osa viikon suosituimmista avoimista julkaisuista oli puolitotuksia. Maanantaina GitHubissa suosituksi noussut työkalu lupaa ajaa valtavia malleja tavallisen pelikoneen näytönohjaimella. Malli käynnistyy, koska työkalu lukee sitä levyltä pala kerrallaan. Laskennallisesti yhden sanan tuottamiseen voi kuitenkin kuluu kymmeniä sekunteja tai enemmän, mikä sulkee keskustelemaan käytön pois. Hugging Facen nousevien mallien listalla oli puolestaan kuvageneraattori, jota on ladattu lähes 2,2 miljoonaa kertaa. Tuntematon julkaisija on poistanut siitä Alibaban alkuperäisen mallin sisältösuodattimet. Yrityksen markkinointikuviin se sopii huonosti. Muokkauksista ei ole riippumatonta tietoa ja suodattimitta malli voi tuottaa myös sisältöä, josta käyttäjä voi joutua vastuuseen. Kolmas nouseva hanke lupaa ohjelmointiavustajat ilmaiseksi ohjaamalla pyynnöt 61 palveluntarjoajan ilmaiskiintiöiden kautta. Silloin jokainen kysymys ja ohjelmarivi kulkee palveluntarjoajalle, jonka tietosuojakäytännöistä käyttäjä ei välttämättä tiedä mitään.

Uusimmat kärkimallit eivät aina toimi heti niissä työkaluissa, joiden kautta niitä käytetään. Riippumattomat vertailut ovat pääosin vahvistaneet Googlen lupaukset uuden Gemini-mallin kyvyistä. Googlen oman komentorivityökalun testiversio ilmoittaa silti, ettei työkalua ole optimoitu yhtiön uusimmalle

mallille. Suositussa avoimen lähdekoodin ohjelmointiavustajassa jokainen pyyntö OpenAI:n uusimmille malleille epäonnistui, kunnes korjaus julkaistiin viikon lopulla. Uutta mallia arvioivan kannattaa kokeilla sitä palveluntarjoajan omassa sovelluksessa ennen kuin vetää johtopäätöksiä sen tasosta.

Arjen avustajille voi antaa yhä isompia ja pidempiä tehtäviä. Tunnettu kehittäjäbloggaaja pyysi Clauden puhelinsovellusta tavallisessa keskustelutilassa rakentamaan selaimessa toimivan pelin, jossa vertaillaan kahta tietokoneen piirtämää puuta. Avustaja työskenteli yhden pyynnön varassa yli kolme tuntia. Kirjoittajan mukaan kehoitus varmistaa, että peli on hauska, toimi epämurkavan hyvin. Avustajalle kannattaa siis kertoa tehtävän lisäksi, millainen lopputuloksen pitää olla.

OpenAI nosti opettajien päivän kunniaksi esiin ChatGPT:n opiskelutilan. Se ohjaa oppilasta ratkaisemaan tehtävän vaihe vaiheelta kysymysten avulla eikä anna valmista vastausta. Kotona ja koulussa tila tarjoaa vaihtoehdon sille, että tekoäly kirjoittaa läksyt oppilaan puolesta. Anthropic puolestaan avasi suurimman Claude-mallinsa nopean tilan kaikille, jotka käyttävät mallia omien sovellustensa kautta. Julkaisutiedoissa ei mainita tilan hintaa, joten se kannattaa selvittää ennen kuin nopeasta tilasta tehdään oletus.

Kolme asiaa kannattaa hoitaa ajoissa. Ensimmäinen koskee Anthropicin erillistä suunnittelupalvelua osoitteessa claude.ai/design, joka sulkeutuu viimeistään 14.12.2026. Tieto tuli esiin yhtiön ohjelmointityökalun päivityksestä. Sulkemisilmoitus ei kerro, siirtyvätkö toiminnot muualle Claudeen. Palvelussa tehdyt suunnitelmat ja prototyypit kannattaa viedä talteen jo nyt eikä jättää joulukuun kiireisiin.

Toinen asia koskee yrityksiä, jotka ovat kesän jälkeen rakentaneet tekoälyavustajalle oman liitännän. Liitäntä on sovitin, jonka kautta avustaja pääsee yrityksen omiin järjestelmiin ja se kannattaa nyt testata uudelleen. Liitäntöjen avoimen standardin virallinen aloitusohje neuvoi heinäkuusta lähtien rakentamaan liitännän tavalla, joka ei toimi standardin uusinta versiota käyttävien avustajien kanssa. Virhe jäi piiloon, koska standardin oma testityökalu käytti oletuksena vanhaa yhteyskäytäntöä. Ohje korjattiin maanantaina.

Kolmas asia koskee työnhakijoita, joiden kannattaa lukea tarkasti, mitä valmispohjat lupaavat. Avoimen lähdekoodin työnhakupohja käyttää Anthropicin ohjelmointityökalua ansioluettelon räätälöintiin, hakemuskirjeisiin ja haastatteluun valmistautumiseen. Tekijä kuvaa pohjaa omalla koneella toimivaksi. Itse tekoälymalli toimii silti Anthropicin palvelimilla, joille ansioluettelo ja muut henkilötiedot lähetetään.

03 Markkina, sääntely ja turvallisuus

Viikon suurin rahaluku tuli Claude-avustajaa kehittävältä Anthropicilta. Financial Timesin mukaan yhtiö kertoi sijoittajilleen, että sen liikevaihto heinäkuusta syyskuuhun oli 16,8 miljardia dollaria eli lähes 50 % enemmän kuin edellisellä vuosineljänneksellä. Vuoden ensimmäisellä neljänneksellä luku oli noin 4,8 miljardia dollaria ja toisella 11,5 miljardia dollaria. Lehden mukaan yhtiö sai liikevaihtoa 1,70 dollaria jokaista laskentatietoa käyttämäänsä dollaria kohden. Sen oikaistu liikevoittoprosentti oli 11 %, kun osakepohjaisia palkkioita ei lasketa mukaan. Tiedot ovat tilintarkastamattomia. Ne perustuvat nimettömiin lähteisiin, jotka kertoivat pörssilistautumista valmistelevan yhtiön sijoittajatilaisuudesta. Sunnuntaina Anthropic julkaisi tiedotteen, jonka otsikko puhuu yleishyödylliseen tarkoitukseen sijoittamisesta ja jonka verkko-osoite viittaa rahastoon.

Avointen mallien markkinassa asetelma muuttui. Kiinalaiset yhtiöt, kuten DeepSeek ja Alibaba, ovat pitäneet avointen mallien kärkeä sen jälkeen, kun Metan edellinen mallisukupolvi jäi keväällä 2025 odotuksista. Riippumaton analyysijulkaisu Interconnects otsikoi lauantaina yhdysvaltalaisen avointen mallien palanneen. Meta kutsuu torstaina julkaistua malliaan kyvykkäimmäksi yhdysvaltalaiseksi avoimeksi malliksi, mutta väite on yhtiön oma eikä riippumaton mittaus. Kiina ei silti kadonnut kärjestä. DeepSeek julkaisi kevyestä mallistaan kuvia ymmärtävän version. Saksalaisen Aleph Alphan saksaa ja englantia osaava malli toimii nyt yhdessä suosituimmista omalla koneella käytettävistä ohjelmista.

Tekoälyinvestointien suunnasta saatiin viikon aikana ristiriitaisia merkkejä. Hollantilainen ASML valmistaa laitteita sirujen tuotantoon ja uutistoimisto Reutersin mukaan sen vahvat tilaukset hälvensivät pelkoa tekoälyinvestointien hiipumisesta. Startupfortune-sivuston mukaan pankit taas yrittävät myydä eteenpäin lainoja, jotka ne vasta järjestivät tekoälykonesalien rakentamiseen. Kansainvälinen valuuttarahasto IMF varoitti brittilehti Guardianin mukaan, että tekoälyosakkeiden mahdollinen romahdus uhkaa koko maailman rahoitusvakautta.

Rahaa riitti silti Modalille, joka vuokraa tekoälyohjelmille eristettyjä ajoympäristöjä. Latent Spacen mukaan yhtiö keräsi rahoitusta 16 miljardin dollarin arvostuksella sen jälkeen, kun sen vuotuinen toistuva liikevaihto ylitti 1 miljardin dollarin. Eristäminen, jonka puutetta tietoturva-asiantuntijat samaan aikaan moittivat, on sijoittajille jo miljardiluokan liiketoimintaa. Tekoälyyhtiö Typesafe AI ilmoitti maanantaina

keränneensä 870 miljoonaa dollaria 7,5 miljardin dollarin arvostuksella.

Suomessa tekoälyn tulevaisuus kytkeytyy yhä tiiviimmin sähköverkkoon. Työ- ja elinkeinoministeriön lakiluonnos asettaisi suuret datakeskukset jonon perälle, kun sähköverkon niukkaa kapasiteettia jaetaan. Tivin mukaan ala syyttää valtiota perustuslain vastaisesta syrjinnästä. Suomi on houkutellut datakeskuksia viileällä ilmastolla ja vakaalla sähköllä. Lakiluonnoksessa sama sähkö on niukkuushyödyke, jonka jakamisesta viranomaiset joutuvat päättämään. EU rakentaa samaan aikaan omia tekoälyn laskentakeskuksiaan eli niin sanottuja tekoälyjättitehtaita. Alan mittausjärjestö MLCommonsin mukaan unionin hankintakilpailussa käytetään sen avointa suorituskyvyn mittausta. Järjestön tiistaina avaamassa uudessa mittauksessa myös sähkönkulutuksen raportointi on pakollista. Pidemmällä aikavälillä sähkön saatavuus vaikuttaa siihen, käsitelläänkö suomalaisten yritysten tekoälypyynnöt Suomessa, muualla Euroopassa vai Euroopan ulkopuolella, millä on suora yhteys tietosuojaan.

Tekoälyverkosto AI Finlandin toimitusjohtaja siirtyy uusiin tehtäviin 16.11.2026. Hän on johtanut verkostoa sen perustamisesta maaliskuussa 2024 lähtien. Sinä aikana verkostoon on liittynyt lähes 500 jäsenorganisaatiota. Verkosto on myös tuonut Kajaanin LUMI-supertietokoneen tekoälytehtaan laskentatietoa suomalaisten toimijoiden ulottuville. Tietotekniikkapalvelujen alalla sopeutus jatkui. Futuricen muutosneuvottelut päättyivät niin, että työt loppuvat 14 henkilöltä. Netum puolestaan aloitti neuvottelut, jotka koskevat 40:tä henkilöä ja voivat johtaa 10–20 tehtävän vähentymiseen tai muuttumiseen. Uutisissa vähennyksiä ei kytketä tekoälyyn.

Pääomasta käytiin viikon aikana kaksi keskustelua. Mistral AI:n perustaja sanoi Tivin mukaan, että eurooppalainen huipputekoäly maksaisi miljardeja ja että vaihtoehto olisi paljon parempi. Mistral on EU:n merkittävin omia suuria tekoälymalleja kehittävä yhtiö. Se on myös yksi harvoista vaihtoehtoista yritykselle, joka haluaa palveluntarjoajansa olevan eurooppalainen. Sunnuntaina Tivi kertoi säätiökentällä avatusta keskustelusta. Sen mukaan suomalaisten säätiöiden noin 30 miljardin euron varallisuudesta voisi löytyä pääomaa, joka nelinkertaistaisi kasvuyritysten rahoituksen.

Toimittajat julkaisivat omia menestyslukujaan runsaasti. OpenAI kertoi analyysissään agenttien käytön kasvavan EU:n yrityksissä ja pienyrityksissä. Arvion on tehnyt yhtiö, joka myy näitä palveluja. Yhtiön asiakastarinan mukaan Stack Overflow'n kaksi insinööriä rakensi muutamassa viikossa palvelun, jonka ChatGPT-liitännäinen kolminkertaisti

rekisteröitymiset. Tietoturvayhtiö SentinelOne kertoo lyhentäneensä haittaohjelmien analysoinnin viikoista tunteihin. Luvut ovat myyjän ja asiakkaiden omia ilmoituksia eivätkä riippumattomia mittauksia. Toimialakohtaisissa hankkeissa yleiskäyttöistä mallia sovitetaan yhden alan työhön alan omien asiantuntijoiden kanssa. Yhdysvaltalainen lakitoimisto Sullivan & Cromwell kehitti yhdessä OpenAI:n kanssa työkalun, joka etsii ongelmakohtia sopimusluonnoksista.

Sääntelyn puolella viikon selvin esimerkki tuli Isosta-Britanniasta. Tutkivaa journalismia julkaisevan openDemocracyn selvityksen mukaan Amazon arvioi tekoälyn avulla ammattiliittoon järjestäytymisen riskiä brittiläisillä varastoillaan. Iso-Britannia ei kuulu EU:n tekoälyasetuksen piiriin. EU:ssa työntekijöiden valintaan, seurantaan ja arviointiin käytettävät tekoälyjärjestelmät luokitellaan suuririskisiksi, jolloin niiltä vaaditaan muun muassa riskienhallintaa, dokumentointia ja ihmisen valvontaa. Työnantajan on myös kerrottava työntekijöille ja heidän edustajilleen ennen tällaisen järjestelmän käyttöönottoa. Suomessa teknisien menetelmin toteutettu työntekijöiden valvonta on jo nyt käsiteltävä yhteistoimintamenettelyssä yksityisyyden suojasta työelämässä annetun lain 21 §:n mukaan. Henkilöstöjärjestelmää hankkivan yrityksen kannattaa kysyä toimittajalta suoraan, mitä järjestelmän tekoälyominaisuudet arvioivat ja millä tiedoilla.

Sama luokitus koskee rekrytointia. Tivin mukaan tekoälyn tekemät työhaastattelut ovat yleistyneet räjähdysmäisesti. Maailmalla levinneessä videossa tubettaja pani tekoälyn vastaamaan tekoälyhaastattelijan kysymyksiin ja haastattelu ajautui kaaokseen. EU:n tekoälyasetus luokittelee rekrytoinnissa käytettävät tekoälyjärjestelmät suuririskisiksi ja asetuksen velvoitteet astuvat voimaan vaiheittain. Microsoftin varapuheenjohtaja kärjisti BBC:n podcastissa, että pesukoneilla ja maitotölkeillä on enemmän sääntöjä kuin tekoälyllä. Hän varoitti alisääntelyn pahoista seurauksista. Ison-Britannian tekoälyn turvallisuusinstituutti julkaisi hahmotelman teknisistä parhaista käytännöistä kehittyneimpien mallien tekijöille. Instituutti kuvasi hahmotelmaa askeleeksi kohti yhteisiä käytäntöjä eikä valmiiksi säännöstenä.

Tavallista yritystä koskevat turvallisuusuutiset liittyivät viikolla enemmän päivityksiin ja huijauksiin kuin tekoälyyn. Kyberturvallisuuskeskus varoitti kahdesta haavoittuvuudesta, joita käytetään jo hyökkäyksissä. Ensimmäinen koskee Ivantin etäyhteyksien ratkaisua, jolla työntekijät yhdistävät koneensa yrityksen verkkoon kotoa tai matkoilta. Toinen koskee Ciscon sähköpostin suojaustuotetta, jonka kautta hyökkääjä voi ajaa laitteella komentoja täysillä ylläpito-oikeuksilla. Tietotekniikasta vastaavalta kumppanilta kannattaa varmistaa, että korjaukset on asennettu ja mahdolliset murren jäljet

tarkistettu, jos yrityksen etäyhteydet tai sähköposti kulkevat näiden tuotteiden kautta.

Kyberturvallisuuskeskuksen viikkokatsauksen mukaan Microsoft 365 -tilien murrot lisääntyvät yhä. Tilin kautta rikollinen pääsee usein sähköpostin lisäksi tiedostoihin, kalentereihin ja sisäisiin keskusteluihin. Tärkein yksittäinen suoja on monivaiheinen tunnistautuminen. Siinä kirjautuminen vaatii salasanan lisäksi esimerkiksi vahvistuksen puhelimesta. Tanskassa nähtiin, mitä heikko salana voi maksaa: yhdeksän miljoonan ihmisen tiedot varastettiin salasanan 123456 takaa. Murren tekijäksi tunnustautunut henkilö antoi paikallismedialle hyökkäyksessä käytetyn ohjelman koodin. Uutisessa lainatun arvion mukaan mikä tahansa uusi tekoälymalli pystyisi vastaamaan.

Huijausten kirjo laajeni. Kyberturvallisuuskeskuksen mukaan rikolliset hyödyntävät nyt majoitusvarausten tietoja. Viestiin, joka viittaa oikeaan varaukseen mutta pyytää maksamaan uudelleen tai vahvistamaan korttitiedot, kannattaa suhtautua epäluuloisesti. Kaakkois-Suomen poliisilaitoksen alueella huijarit saivat edeltävällä viikolla yli 80 000 euroa. Poliisi varoittaa erityisesti videopuheluista, joissa ruudulla näkyy virkapukuinen henkilö. Epäilyttävän puhelun voi aina katkaista ja soittaa viranomaiselle itse tämän virallisesta numerosta. Turvatilihuijauksissa suomalaiset ovat menettäneet tänä vuonna selvästi vähemmän rahaa kuin aiemmin. Rikolliset yrittävät nyt kiristää rahaa arkaluonteisten videoiden varjolla. Viestit on rakennettu herättämään voimakas tunnereaktio.

Tekoäly näkyy turvallisuudessa enemmän tulevaisuuden riskinä kuin tämän päivän murtoina. Tutkimuslaitos Epoch AI arvioi perjantaina, että avoimet mallit ovat kyberhyökkäyksissä tarvittavissa taidoissa enää kuusi tai seitsemän kuukautta suljettujen kärkeimallien jäljessä. Avointen mallien kärjessä on kiinalaisen Z.ai-yhtiön malli. Haavoittuvuuksien etsimiseen kelpaavat taidot leviävät siis noin puolen vuoden viiveellä kenen tahansa ladattaviksi. Päivitykset ja monivaiheinen tunnistautuminen ovat siksi entistä tärkeämpiä. Ethereum-säätiön tutkija kehotti kryptoalaa siirtymään niin sanottuun bunkkeritilaan. Hänen mukaansa tekoälyn nopea kehitys matematiikassa voisi murtaa lompakoiden suojauksen kuukausien kuluessa. Varoitus on ennuste eikä raportti tapahtuneesta murrosta.

Tekoälytyökaluissa itsessään havaittiin myös puutteita. Anthropicin virallisessa lisäosakokoelmassa on tietoturvatarkastaja, joka käy ohjelmoijan koodimuutokset läpi automaattisesti. Se jätti tietyissä tilanteissa osan koodista lukematta. Korjausten jälkeen epäonnistunut tarkastusta ei enää ilmoiteta puhtaaksi. Yrittäjä, joka teettää ohjelmistoja tekoälyn avulla, voi kysyä kehittäjiltään, onko työkalu päivitetty ja ajettiin aiemmat tarkastukset uudelleen.

04 Mitä seurata

Metan avoimen mallin ehdot EU:ssa. Metan aiempien avoimien mallien käyttöehdot ovat rajanneet EU:hun sijoittautuneiden yritysten oikeutta käyttää kuvia ymmärtäviä malleja. Torstaina julkaistun mallin ehdot ratkaisevat, kuinka moni suomalainen yritys voi ottaa sen käyttöön sellaisenaan. Malli ehti parissa päivässä muiden yritysten mallipalveluihin. Seurattava merkki on se, tarjoavatko eurooppalaiset palveluntarjoajat mallia EU:ssa sijaitsevilta palvelimilta.

Ulkopuolinen tarkastus tekoälymalleille. Yhdysvaltalainen tutkimusjärjestö METR arvioi tekoälymallien riskejä. Se julkaisi viisi kirjoitusta siitä, miten ulkopuolinen arvioija voisi tutkia julkaisemattomia malleja. Seuraavana päivänä Anthropic julkaisi päivituksen riippumattomista arvioijista. Samalla viikolla Anthropicin tutkimus esitti, että mallit saattavat kyetä kopioimaan itsensä ja toimimaan ilman suurta konesalia. Toinen tutkimusartikkeli väittää todistavansa matemaattisesti, että malli voi kätkeä minkä tahansa kykynsä. Molemmat koskevat samaa kysymystä: kertooko ulkopuolinen testaus luotettavasti, mihin malli pystyy. EU:n tekoälyasetus velvoittaa suurimpien yleiskäyttöisten mallien tarjoajat arvioimaan ja testaamaan malliensa riskejä. Kehityksen suunnan näkee siitä, alkavatko uusien mallien julkistuksiin liittyä ulkopuolisten tarkastajien raportit samaan tapaan kuin tilinpäätökseen kuuluu tilintarkastajan lausunto.

Puolen vuoden viive. Epoch AI:n mittaama ero avointen ja suljettujen mallien kybertaidoissa on luku, jonka muutosta kannattaa seurata. Nykyisellä viiveellä suljettujen kärkimallien tämän syksyn kybertaidot olisivat kenen tahansa ladattavissa ensi kevääseen mennessä. Puolustajat saavat vähemmän aikaa päivitysten asentamiseen, jos ero kapenee.

Virheraporttien ruuhka. Tietokantahankkeen viivästys voi jäädä yksittäistapaukseksi. Se voi myös olla ensimmäinen

merkki siitä, että ohjelmistojen julkaisutahti alkaa riippua ihmisten käsittelykyvystä. Suunnan näkee siitä, alkavatko muut avoimen lähdekoodin hankkeet rajata koneellisesti tuotettuja virheraportteja tai siirtää julkaisujaan. Yritysten käyttämissä ohjelmistoissa sama voi näkyä harvempina mutta perusteellisemmin tarkistettuina päivityksinä.

Valmiin työn hinta. Alan mittausjärjestö MLCommons sai Yhdysvaltain kansalliselta tiedesäätiöltä 3,5 miljoonan dollarin rahoituksen kahdeksi vuodeksi. Rahalla rakennetaan yhteinen mittari sille, kuinka paljon laskentaa ja energiaa agenttien kokonaiset tehtäväketjut kuluttavat. Valmis mittari antaisi ostajalle tavan verrata palveluja sen mukaan, mitä valmiiksi saatu työ maksaa. Päätösmallien yleistymisen näkee siitä, että palvelut alkavat hinnoitella nopeat valinnat erikseen, esimerkiksi asiakasviestien ohjaamisen oikealle käsittelijälle.

Datakeskuslaki. Lakiluonnoksen jatkokäsittely ja työ- ja elinkeinoministeriön vastaus perustuslakikritiikkiin kertovat, rakennetaanko Suomeen lähivuosina suuria tekoälyn laskentakeskuksia vai siirtyykö kapasiteetti muualle.

Anthropicin listautuminen ja Character.ai:n kauppa.

Anthropicin tilintarkastetut luvut tulevat julki vasta listautumisasiakirjoissa. Niistä selviää, kestäkö Financial Timesin kertoma kasvu tilintarkastajan tarkastelun. Character.ai:n saneerauksessa seurattavaa on, kuka palvelun lopulta ostaa ja mitä ostajan ehdoissa luvataan käyttäjien keskusteluista.

Kalenteriin. AI Finlandin toimitusjohtaja vaihtuu 16.11.2026 ja uuden toimitusjohtajan haku alkaa lähiaikoina. Anthropicin erillinen suunnittelupalvelu sulkeutuu viimeistään 14.12.2026. Netumin uusi organisaatio tulee voimaan viimeistään 1.1.2027.

OSA 2 • TEKNISILLE LUKIJOILLE

Tekninen osa

Teknisille lukijoille: mallit, alustat, työkalut, vertailut ja tietoturva yksityiskohtineen. Osa 1 kertoi viikon olennaisen arkikielellä.

TÄSSÄ OSASSA

- | | | | |
|----|-------------------------------|----|-------------------------------|
| 05 | Suuret toimijat | 09 | Työskentelytavat ja agentit |
| 06 | Avoimet mallit ja ekosysteemi | 10 | Liitännäiset, MCP ja repot |
| 07 | Ajoalustat ja infra | 11 | Tietoturva ja haavoittuvuudet |
| 08 | Benchmarkit ja vertailut | | |

05 Suuret toimijat

Viikon painavimmat mallijulkistukset tulivat Googelta ja Metalta. Kummassakin tapauksessa riippumaton mittaus ehti muutamassa päivässä joko vahvistaa yhtiön omat väitteet tai suhteuttaa niitä.

Google DeepMind julkaisi tiistaina 13.10. Gemini 4 Argonin ja samana päivänä erillisen Gemini 4 Argon for Robotics -julkistuksen. Gemini Robotics -mallit ovat aiemmin tulleet perusmallin jälkeen omana linjanaan. Argon siirtyi heti yleiseen saatavuuteen Gemini API:ssa ja Googlen 100 dollarin kuukausitilauksessa. Rajapinnassa miljoona syötetokenia maksaa 2 dollaria ja miljoona tulostetokenia 10 dollaria. Latent Spacen AINewsin mukaan useimmat ulkopuoliset vertailut, CursorBench mukaan lukien, vahvistavat Googlen omat väitteet riippumattomasti. Komentorivityökalun llm-gemini versio 0.35 tukee mallin päättelytasoja low, medium ja high. High-tasolla piirretty SVG-pelikaani maksoi 17,166 senttiä. DeepMind julkaisi seuraavana päivänä kirjoituksen Argonin vastuullisesta ja turvallisesta kehittämisestä.

Googlen omat työkalut eivät ole yhtä valmiita kuin malli. Gemini CLI:n loppuviikon nightly-versio lisäsi bannerin, jonka mukaan CLI:tä ei ole optimoitu Gemini 4:lle (#29774). Rajapinnan muutoslokissa Flash-linjan uusien merkintä on edelleen gemini-3.8-flash.

Metan viikko eteni kahdessa vaiheessa. Meta Superintelligence Labs julkaisi maanantaina 12.10. Muse Wonderin ja Muse Wonder Fullin, joita Meta kutsuu kyvykkäimmiksi malleikseen. Artificial Analysisin Intelligence Indexissä ne saivat Max-asetuksella 52 ja 56 pistettä. Artificial Analysis pitää Muse Wonderin Max-tasoa GPT-6.1 Solin veroisena ja Latent Space luokittelee mallit Claude 5.5 -tasoisiksi. Päättelytasoja on kuusi minimalista maxiin. Julkaisuväli on pysynyt kuukauden mittaisena: Muse 1.2 ilmestyi elokuun 5. päivänä, Muse 1.3 syyskuun 2. päivänä ja Wonder lokakuun 12. päivänä.

Kumpikaan maanantain lähde ei kertonut, julkaiseeko Meta Wonderin painoja. Avoimen puolen vastaus tuli torstaina 15.10. toisena mallina. Muse Spark Open on avoimin painoin julkaistu multimodaalinen asiantuntijaseosmalli. NVIDIA:n teknisen blogin mukaan siinä on 816 miljardia parametria ja miljoonan tokenin konteksti. Tokenia kohden aktiivisena on 42 miljardia parametria. NVIDIA esittelee mallin agenttisen koodauksen kuormana GB300 NVL72 -räkissä. Pelkät FP8-painot vievät noin 816 gigatavua, joten nelibittinenkään kvantisointi ei saa mallia mahtumaan yhdelle kiihdyttimelle.

Meta kutsuu Spark Openia kyvykkäimmäksi yhdysvaltalaiseksi avoimen painon malliksi. Interconnectsin lauantainen analyysi tulkitsee yhtiön liikkeen muutokseksi sekä avoimessa että suljetussa kärjessä ja yhdysvaltalaisen avointen painojen paluuksi. Llama 4:n jälkeen avoimen painon kärkeä ovat pitäneet lähinnä Qwen, DeepSeek, Moonshot ja Zhipu. Intelligence Indexissä Spark Open sai xhigh-asetuksella 39 pistettä. Metan suljetun ja avoimen kärkimallin väliin jää siten 17 pisteen ero, tosin eri päättelyasetuksilla mitattuna. Malli otettiin nopeasti käyttöön. Cline asetti sen kahdessa päivässä Eden AI:n ja Prime Inferencen oletusmalliksi.

Anthropic julkaisi 12.10. tutkimuksen AI models may be capable of self-replicating without access to a datacenter (Hacker News 95 pistettä, 107 kommenttia). Reuters muotoili saman seuraavana päivänä ennusteeksi otsikolla AI models may soon self-replicate without access to data centers. Tutkimuksen otsikko puhuu kyvystä ja uutistoimiston otsikko ajankohdasta. Tuotepuolella Claude Opus 5.5:n fast mode poistui 15.10. tutkimusesikatselusta. Pääsyä ei enää tarvitse pyytää eikä fast-mode-2026-02-01-beetaotsaketta tarvita. Otsaketta edelleen lähettävät pyynnöt toimivat, joten olemassa oleviin integraatioihin ei tarvitse koskea. Claude Coden /design-sync- ja /design-login-komennot ilmoittavat nyt erillisen Claude Design -palvelun sulkeutuvan viimeistään 14.12.2026.

OpenAI:n viikko koostui jakelusta ja asiakastarinoista, joiden luvut ovat asiakkaiden ja OpenAI:n omia. Third Moment Research kertoo ajavansa 10 000 agenttia yli 36 tunnin jaksoissa GPT-6 Astralla, GPT-6.1 Solilla, Codexilla ja Daybreak-tietoturvatutoteella. Stack Overflow for Agents toteutettiin kahden insinöörin voimin Codexilla muutamassa viikossa. Käytännön kannalta hyödyllisin tieto ei tullut tiedotteista vaan rajapinnasta. GPT-6-perheen tunnuksset, kuten gpt-6-sol ja gpt-6.1-sol, vaativat OpenAI- ja Azure-päätepisteissä max_completion_tokens-parametrin. Clinen versiot ennen CLI 3.0.73:a, Desktop 0.0.48:aa ja SDK 0.0.95:tä lähettivät max_tokens-parametrin, jolloin jokainen pyyntö hylättiin. Korjattu SDK käyttää uutta parametria tunnuksille gpt-5:stä gpt-9:ään ja pitää Azuren gpt-35-turbolla vanhan. Omat OpenAI-yhteensopivat asiakkaat kannattaa tarkistaa samasta syystä.

OpenAI ei julkaissut viikolla omaa tiedotetta päätösrapiapinnastaan (decisions endpoint). Tieto rajapinnasta perustuu Opperin Show HN -julkaisuun. Rajapinnan ympärille syntynyt päätösmallien luokka nousi silti viikon aikana ajoalustojen keskeiseksi teemaksi.

Mistral julkaisi Timestralin, todennäköisyysennusteisiin tarkoitettun aikasarjamalliperheen. Kokoja on neljä ja parametreja on viidestä miljoonasta 209 miljoonaan. Yhtiö lupaa tehokkaan ajon suorittimella ja zero-shot-käytön ilman hienosäätöä. Kilpailijoita ovat Amazonin Chronos, Googlen TimesFM ja Salesforcen Moirai. GIFT-Evalin kaltaisissa riippumattomissa aikasarjavertailuissa Timestralista ei vielä ole tuloksia. Painojen saatavuus ja lisenssi on varmistettava Mistralin julkistuksesta. The Batch käsitteli viikolla myös

Mistral Large 4:ää. xAI julkaisi X Subscription -tilauksen ja Grok Voice 3.0:n.

Useampi suuri toimija suuntasi julkistuksiaan biotieteisiin. Google DeepMind kertoo nopeuttaneensa antimikrobisten peptidien etsintää Antigravity for Science -ympäristöllä. Google Research kuvasi autonomisen märklaboratorion rakennuspalikoita. OpenAI:n Rosalind Academic Plan antaa kelpoisille akateemisille laboratorioille sponsoroidun pääsyn GPT-Rosalindiin ja sen työkaluihin.

06 Avoimet mallit ja ekosysteemi

Avointen mallien viikon voi lukea kolmen luvun kautta. Kokonaisparametrit kertovat mallin koon ja tokenia kohden aktiiviset parametrit kertovat dekodauksen nopeuden. Muistin tarve, johon kaikkien painojen on mahduttava, ratkaisee, millä koneella malli ylipäätään käynnistyy.

Muse Spark Openin rinnalle kärkipäähän tuli Reflection AI:n Beam 1.0. Yhtiö keräsi vuonna 2025 noin 2 miljardin dollarin rahoituksen nimenomaan avointen kärkimallien rakentamiseen Yhdysvalloissa. Beam on sen ensimmäinen julkaistu malli, jonka Hugging Facen blogi esittelee 501 miljardin parametrin avoimena mallina. Painot vievät 16-bittisinä noin teratavun ja kahdeksanbittisinä noin puoli teratavua. Malli vaatii siis usean kiihdyttimen palvelimen, eikä se toimi työasemalla. Intelligence Indexissä Beam sai 30 pistettä. Laboratorio on perustettu nimenomaan avoimen kärjen rakentamiseen, mutta sen ensimmäinen malli jäi yhdeksän pistettä Metan avoimen mallin taakse.

NVIDIA julkaisi Nemotron 3.5 -perheeseen kaksi mallia jatkuvasti käynnissä oleville yritysagenteille. Super VL rakentuu 120 miljardin parametrin hybridille MoE-rungolle, jossa on sekä Mamba-2- että Transformer-kerroksia. Tokenia kohden aktiivisena on 12 miljardia parametria. Kuvia malli käsittelee C-RADIO-kooderilla. Mukana ovat myös työkalukutsut, päälle ja pois kytkettävä ajattelutila ja miljoonan tokenin konteksti. Superista tislattu Bolt on 74 miljardin parametrin malli, jossa aktiivisena on yhdeksän miljardia. NVIDIAn mukaan se on viritetty 24 gigatavun näytönohjaimelle ja mukana tulee luonnosmalli spekulatiiviseen dekodaukseen. Nelibittisenä 74 miljardia parametria vie kuitenkin noin 37 gigatavua. Bittimäärää on laskettava tai osa asiantuntijoista on pidettävä keskusmuistissa, jotta malli mahtuisi 24 gigatavuun. Ollama toi molemmat mallit käyttöön julkaisupäivänä komennolla `ollama run nemotron-3.5-super`. Intelligence Indexissä Super VL sai 26 ja Bolt 21 pistettä.

Eurooppalaisista malleista arkkitehtuuriltaan kiinnostavin on Aleph Alphan Kolibri-1, jolle llama.cpp toi tuen buildissa

b11709. Saksaa ja englantia käsittelevässä MoE-päätelmämallissa on 78 miljardia parametria, joista 3,46 miljardia on aktiivisena tokenia kohden. Kerroksia on 50, reititettyjä asiantuntijoita 384 ja jaettuja asiantuntijoita yksi. Huomiokerrokset vuorottelevat suhteessa neljä yhteen. Neljää liukuvan ikkunan kerrosta (ikkuna 513 tokenia) seuraa yksi täysi kerros. Liukuvat kerrokset käyttävät NeoX-tyylistä RoPEa, kun taas täydet kerrokset toimivat kokonaan ilman paikkakoodausta (NoPE). GQA:ssa 48 kyselypäättäjää jakaa neljä KV-päätä. KV-välimuisti jää siksi pieneksi. Täydet kerrokset vievät fp16-muodossa noin 20 KiB tokenia kohden ja liukuvien kerrosten osuus rajautuu 513 tokeniin. Noin 4,5 bitin kvantisoinnilla painot vievät karkeasti 44 Gt. Tokenia kohden laskentaa on silti vain noin 3,5 miljardin parametrin tiheän mallin verran, mikä sopii yhtenäismuistikoneille ja suoritinaajaan. Lisenssiehdot eivät käy ilmi llama.cpp:n muutoksesta. Toinen eurooppalainen julkaisu on Swiss AI -aloitteen Apertus 1.5. Sarjan takana ovat EPFL, ETH Zürich ja Sveitsin kansallinen supertietokonekeskus CSCS, jotka julkaisevat painojen lisäksi koulutusdatan ja -reseptit.

Pienemmistä malleista viikolla huomiota saivat laitteella ajettavat mallit. Liquid AI:n LFM2.5-24B-A2B:ssä on 24 miljardia parametria, joista noin kaksi miljardia on aktiivisena. Yhtiö markkinoi mallia nopeille ja yksityisille agenteille, jotka toimivat käyttäjän omalla laitteella. Nelibittisenä pelkät painot vievät noin 12 Gt, joten 16 gigatavun koneessa jää vähän tilaa kontekstille ja käyttöjärjestelmälle. Malli ladattiin parissa päivässä 4 067 kertaa. OpenBMB julkaisi MiniCPM-V 4.7:n kahtena versiona. Suurempi on 35 miljardin parametrin MoE-malli, jossa noin kolme miljardia parametria on aktiivisena. Pienempi on miljardin parametrin malli, jolle llama.cpp toi tuen torstaina.

Upotusmalleissa viikko oli poikkeuksellisen vilkas. Microsoftin bitnet.cpp julkaisi BitNet-embedding-0.6B:n ja -270M:n, jotka ovat repon mukaan ensimmäiset yksibittiset upotusmallit. Esitäyttö on kahdeksalla säikeellä F16:een verrattuna 1,42–2,28 kertaa nopeampi suuremmalla ja 1,32–

1,74 kertaa nopeampi pienemmällä mallilla. Llama.cpp v0.7.0:n julkaisutiedotteen mukaan Googlen EmbeddingGemma 2 upottaa tekstiä, kuvaa ja ääntä. Hakujärjestelmissä yksi malli kattaa siis kolme modaaliteettia. Hacker Newsin suosituimpiin nousi EG2-projekti (321 pistettä), jossa saman mallin pohjalta tehtiin puheentunnistin. Kuvapuolella Black Forest Labsin FLUX.3 Image Klein sai Diffusers-tuen. Trendilistalla sen ComfyUI-paketti oli ladattu 73 767 kertaa ja perusversio 22 790 kertaa. Malli otetaan siis käyttöön ensisijaisesti ComfyUI-työnkulkujen kautta. Black Forest Labsin malliperheiden lisenssi on aiemmin vaihdellut mallin koon mukaan, joten mallikortti kannattaa lukea ennen kaupallista käyttöä.

DeepSeek-V4.1-Flash-1017 nousi lauantaina trendilistalle 817 tykkäyksellä, vaikka latauslaskuri näytti vielä nollaa. Päiväyspäätte noudattaa DeepSeekin tapaa julkaista samaan mallilinjaan päivitettyjä checkpointeja, kuten aiemmin V3-0324. Kyse on päivitetystä versiosta eikä uudesta mallilinjasta. Clinen lauantainen katalogipäivitys kertoo, miten nopeasti uudet avoimet mallit leviävät palveluntarjoajille. Beam 1.0 on nyt CoreWeaven oletusmalli, Nemotron 3.5 Super VL Pioneerin ja DeepSeek V4.1 Flash Scalewayn.

Latauslistojen kärjessä tilanne näytti toiselta. Maanantain ladatuin trendimalli oli tuntemattoman julkaisijan kvantisoina Qwen-Image-johdannainen, josta sisältösuodatus oli poistettu. Sillä oli 2,19 miljoonaa latausta. Toisen noston nimessä luetellaan viisi opettajamalleiksi tulkittavaa mallia ja neljä menetelmää, mutta arviointituloksia ei ole lainkaan. Malli on tislattu myös suljettujen mallien tuotoksista, jos nimi kuvaa sisältöä. Tällainen tislauksen on ristiriidassa monien kaupallisten palvelujen käyttöehtojen kanssa. Latausmäärä kertoo kysynnästä mutta ei mallin alkuperästä.

Suomen kannalta viikon kiinnostavimmat tutkimukset koskivat tokenisointia ja niukkaa koulutusdataa. Nature-lehdessä julkaistu Retrofitting language models to operate over bytes (2026) käsittelee valmiiden kielimallien muuntamista käsittelemään tavuja tokenien sijaan. Alisanatokenisoijat pilkkovat taivutusrikkaat kielet useampiin tokeneihin kuin englannin, mikä nostaa hintaa ja kuluttaa konteksti-ikkunaa. Tavutaso tasoittaa tätä epäsuhtaa, mutta sekvenssit pitenevät. NVIDIA kuvasi Hugging Facen blogissa intialaisten kielten esikoulutuksen skaalaamista 2 800 miljardiin tokeniin ja avasi datan generointiputken. Menetelmä sopii myös pienille kielille, joista verkossa ei ole riittävästi luonnollista tekstiä.

07 Ajoalustat ja infra

Pyytääkö asiakas hyvitystä? Ollama 0.40.4:ssä kysymys välitetään komennolla `ollama run nimble --ask "Is a refund being requested?"` mallille, joka vastaa pelkällä kyllä tai ei. Samalla mekanismilla tekstiä voi luokitella annettuihin luokkiin tai pisteyttää arviointikriteerejä vasten. Viikon alussa päätösmalleista oli vain toisen käden tietoa. Opper kertoi Show HN -julkaisussaan, että OpenAI on julkaissut päätösrajapinnan ja Cloudflare on tuonut markkinoille clef-mallinsa. Malleja yhdistää pieni viive. Päätökset on tarkoitus tehdä niin nopeasti, että niillä voi ohjata reaaliaikaista järjestelmää.

Viikon aikana luokka levisi kaikkiin keskeisiin avoimiin ajoalustoihin. Maanantaina llama.cpp sai tuen pplx-decider-malleille (b11571 ja b11572). Ollama toi kev-päätösmallit Apple siliconille MLX-ajurin kautta. Torstaina julkaistun llama.cpp v0.7.0:n uusien mallien listalla olivat LiquidAI:n d1 ja d1-omni, vLLM:n decision 2 ja d3 sekä pplx-decider. Samana päivänä build b11642 lopetti päätös-, rerank-, upotus- ja infill-slottien tallentamisen kehotevälimuistiin, josta niitä ei koskaan luettu takaisin.

IBM:n Granite Switch 4.1 tarjoaa keskustelun, päätökset ja puheen samasta pääte pisteestä. Sen decision-tila tuli llama.cpp:hen perjantaina (b11700). Lauantainen muunninkorjaus (b11707) paljastaa, miten tila toimii. Switch-

malli kuvaa jokaisen vaihtoehdon yhdeksi tokeniksi ja tokenien logiteista muodostetaan päätöksen todennäköisyysjakauma. Isoilla ja pienillä kirjaimilla sekä numeroilla saatiin vain 66 erillistä yhden tokenin merkkiä, vaikka malli tukee 255 vaihtoehtoa. Muunnin etsii nyt jopa 256 tällaista merkkiä mallikohtaisesti, koska yhden tokenin merkkijohdistelmät riippuvat tokenisaattorista.

Sunnuntaina vLLM 0.32.0 teki päätösmalleista vakio-ominaisuuden. Uusi `POST /v1/systemone` -pääte piste vastaa rakenteisesta tilasta esitettyihin kolmentyyppisiin kysymyksiin: choice, noul ja score. Pääte piste on käytössä oletuksena ja toimii myös Cloudflaren clef-omnilla, LiquidAI:n d1-3B:llä ja DiffusionGemmalla. Ollama puolestaan toi pilvipalveluunsa clef-mallin ja lisäsi 0.40.7:n julkaisukandidaattiin PPLX Decider v1.2:n tuen MLX-taustaan. Viikossa päätösmallit etenivät Show HN -maininnasta vLLM:n oletuspääte pisteeksi. Versioon 0.32.0 kertyi 814 commitia 363 kehittäjältä, joista 121 oli uusia. Toinen arkea helpottava uudistus koskee jäsentimiä. Valitsimet `--tool-call-parser hf` ja `--reasoning-parser hf` lukevat checkpointin oman `response_template`-kentän, joten mallit, joissa kenttä on mukana, eivät tarvitse mallikohtaista jäsentintä.

Toinen viikon kehityslinja on asiantuntijaseosmallien ajaminen kortilla, jonka muisti on mallia pienempi. Llama.cpp v0.7.0 (build b11669) toi valitsimen `--moe-cache-mib`, joka varaa näytönohjaimelta välimuistin keskusmuistissa pidettäville asiantuntijoille. Projektin mukaan dekooodaus nopeutuu jopa noin 2,2-kertaisesti, kun asiantuntijat on siirretty pois näytönohjaimelta. Ominaisuus sopii juuri Boltin kaltaisille malleille, jotka eivät mahdu kokonaan 24 gigatavuun. Ollaman mukana tuleva llama.cpp oli tiistaina 71 buildia kehityshaaraa jäljessä. Torstain Ollama 0.40.5 perustui jo v0.7.0:aan. Lauantaina CUDA sai grouped experts top-k -fuusion (b11710) MoE-malleille, jotka käyttävät ryhmärajattua reititystä.

Äärimmäisen ratkaisun tarjoaa maanantaina GitHubin trendilistalle noussut lyogavin/airllm. Se lupaa 70 miljardin parametrin mallin päättelyn neljän gigatavun näytönohjaimella lataamalla levyltä vain kulloinkin laskettavan kerroksen. Muistilupaus voi pitää paikkansa, mutta hinta maksetaan levyn kaistassa. 70 miljardia parametria vie 16-bittisenä noin 140 gigatavua. Nopealla NVMe-levyllä (noin seitsemän gigatavua sekunnissa) näiden lukeminen kestää jokaista generoitua tokenia kohden noin 20 sekuntia. Esitöytössä yksi kerroksen lataus palvelee koko syötettä, joten pitkien syötteiden eräajo on siedettävämpää kuin interaktiivinen käyttö. Repo sopii eräajoon, jos mallia ei muuten saa ajettua. Interaktiiviseen käyttöön se ei sovellu. Laitteistopuolella Intelin D270-datakeskusnäytönohjaimeen luvataan jopa 480 Gt LPDDR5X-muistia. LPDDR5X tuo kapasiteettia halvemmalla kuin HBM, mutta kaistanleveys on pienempi. Kortti sopii siksi suurille MoE-malleille, joissa muistin määrä on tärkeämpi kuin kaista tokenia kohden.

Kolmas kehityslinja koskee hybridimallien tilaa. NVIDIA:n ReplaySSM-kirjoitus kuvasi ongelman maanantaina. Huomion KV-välimuistin voi kelata taaksepäin katkaisemalla sen, mutta tila-avaruusmallin tila kirjoitetaan yli jokaisella askeleella. Siksi spekulatiivisen dekooodauksen hylkäykset ja etuliitevälimuisti edellyttävät, että tila tallennetaan tarkistuspisteisiin tai lasketaan uudelleen. NVIDIA lukee hybridimalleihin Nemotron 3:n, Qwen3.5–Qwen3.8:n, Kimi-K3:n ja GLM-5.3-Flashin. Viikon korjaukset kohdistuivat juuri tähän. Ollaman MLX-tausta tiivistää nyt rekurrenttien tilan spekulatiivisen peruutuksen jälkeen. Nemotron-tuen yhteydessä Mamban softplus-funktio vakautettiin. vLLM 0.32.0rc3 korjasi Mamba2:n chunk scan -kernelin NaN-tulosteet Hopper-sukupolven sm90-piireillä.

Spekulatiivisessa dekooodauksessa merkittävin korjaus on llama.cpp:n b11644. Tavallisen KV-välimuistin luonnoskonteksti eli MTP-pää tai tiheä `-md`-luonnosmalli kirjoitti koko KV-välimuistinsa jokaiseen palvelimen tarkistuspisteeseen. Tarkistuspiste kasvoi siksi jokaisella tokenilla luonnosmallin KV-välimuistin verran. Spekulatiivista

dekooodausta palvelimella ajavien kannattaa päivittää vähintään tähän buildiin. MTP-tuki laajeni viikon aikana GLM5-Nextiin ja JetBrainsin Mellum2.1-koodimalliin (b11695). Lisäksi v0.7.0 toi uuden spekulatiivisen dekooodauksen tekniikan, LiliCorrin. Ollama 0.40.7:n julkaisukandidaatti antaa llama-serverin päätellä luonnosmallin spekulatiotyypin itse.

Matalabittinen kvantisointi eteni usealla taustalla. Ollama 0.40.6 lisäsi GGUF-muotoon q2_0-tensori- ja tiedostotyyppit. Llama.cpp:n HIP-tausta purkaa nyt q1_0:n `__builtin_amdgcn_perm`-käskyllä (b11715) ja CUDA:n MMVQ nopeutui matalabittisillä tyypeillä (b11713). Qualcommin Hexagon-taustan HMX-matriisipolku sai sunnuntaina tuen Q1_0:lle ja Q2_0:lle sekä Hadamard-operaatioille (b11716 ja b11717). Laitetuki laajeni myös Intel Arc Pro B65:n ja B70:n shader-määrillä (b11614) ja Intelin ESIMD-ydinten palauttamisella FPN-polulle (b11718).

Hyvä esimerkki huolellisesta mittaamisesta on keskiviikon b11618. Se ottaa RDNA3.5:n HIP-polulla 64-riviset MMQ-tiilet käyttöön niissä tyyppin ja J:n yhdistelmissä, joissa ne todettiin nopeammiksi AMD:n Strix Halo -piirillä (gfx1151). Pienempi tiili jättää LDS-muistia toiselle työryhmälle samassa laskentayksikössä. Testi kattoi 100 MUL_MAT-muotoa kolmella kierroksella. Mediaani parani 2,1 % ja paras tapaus 26,0 %. Arvo J=96 jätettiin tarkoituksella pois, koska se ei hyötynyt muutoksesta.

Yritysverkon välityspalvelin esti latauksia kahdella alustalla ja molemmat korjasivat vian saman viikon aikana. Llama.cpp:n `common_http_client()` ei aiemmin huomionnut `HTTP_PROXY`- ja `HTTPS_PROXY`-muuttujia, joten `-hf`-lataus epäonnistui välityspalvelimen takana ilman virheilmoitusta. Build b11699 lukee nyt https-osoitteille `HTTPS_PROXY`- tai `https_proxy`-muuttujan ja http-osoitteille vastaavat muuttujat. Virheellisesti muotoillut välityspalvelinosoitteet ohitetaan edelleen hiljaa, mikä kannattaa muistaa vianetsinnässä. Ollama 0.40.6-rc1 lisäsi välityspalvelintuen rekisterin uudelleenohjauskohteille. Lopullisen 0.40.6:n muutosluettelo ulottuu kuitenkin vain rc0:aan asti. Korjauksen mukanaolo kannattaa siksi varmistaa ennen kuin latausongelmat todetaan ratkaistuiksi. SYCL:n FP32- ja FP16-buildeihin otettiin käyttöön OpenSSL (b11610), mikä palautti HTTPS-lataukset Hugging Facesta. Lauantaina llama.cpp:n web-käyttöliittymään tuli malliselain, jolla buildista b11714 alkaen malleja voi selata ja ladata suoraan Hugging Face Hubista. Q4_0-mtp-tyyppinen tagi hakee sivutiedoston, jos varsinaista mallitiedostoa ei löydy. Näin pelkän MTP-luonnosmallin tai mmproj-projektorin voi ladata erikseen.

Automaatiokehysten korjaukset koskivat pitkiä ja keskeytyviä ajoja. LangGraph 1.2.15 ja sen checkpoint-kirjastot korjasivat DeltaChannel-tallennusta. Jokainen messages-kanavaan kirjoitettu viesti saa nyt tunnisteiden. Välimuistista palautetun

solmun kirjoitukset tallennetaan samoin kuin ajetun solmun. `interrupt_before`-katkaisukohta pysyy voimassa myös silloin, kun tyhjennystä pyydetään ennen sitä. Versio 1.2.16 korjasi resume-semantiikkaa neljässä kohdassa. Lisäksi checkpoint-postgres 3.2.0 toi PostgresSaveriin ja AsyncPostgresSaveriin prune-toiminnon, jolla tarkistuspistetaulujen kasvua voi hallita. Interruptiin ja `Command(resume=...)`-jatkoon perustuvat ihminen silmukassa -työnkulut kannattaa testata päivityksen jälkeen. n8n päivitti kolmea julkaisulinjaa rinnakkain. Käytännössä tärkein muutos on 2.43.7:ssä. Se keskeyttää kesken olevat tuotantoajot hallitusti, kun worker sammutetaan, mikä helpottaa jonotilassa toimivien asennusten päivittämistä ja skaalaamista.

Open WebUI 0.12.1 toi kanaviin ulkoiset sovellukset. Ylläpitäjä liittää sovelluksen HTTPS-osoitteella ja tarkistaa sen pyytämät oikeudet. Sovellus näkyy kanavan jäsenlistassa ja vastaa @-mainintoihin samaan tapaan kuin yrityspikaviestimien botit. NVIDIA avasi Run:ai-alustan koko lähdekoodin osana DSX AI Factory -kokonaisuutta. Kubernetes-pohjainen GPU-kuormien ajoitus- ja orkestrointikerros tarjoaa omaa klusteria ylläpitäville tavan hoitaa GPU-jako ja jonotus ilman erillistä lisenssiä. GitHubin trendilistalle nousi LightSeekin TokenSpeed, agenttikuumille

suunniteltu päättelymoottori. Sen ajastimessa on C++-ohjaustaso ja Python-suoritusaso. Pyyntöjen elinkaari, KV-välimuistin omistajuus ja limityksen ajoitus on mallinnettu äärellisenä tilakoneena. KV-resurssien turvallinen uudelleenkäyttö varmistetaan tyyppijärjestelmällä käännösaikana. Projekti lupaa TensorRT-LLM:n suorituskyvyn ja vLLM:n käytettävyyden, mutta riippumattomia mittauksia ei vielä ole.

Agenttien hiekkalaatikoista on tullut kasvava kuorma infrastruktuurille. Latent Spacen mukaan Modal keräsi D-rahoituskerroksen 16 miljardin dollarin arvostuksella ja ylitti miljardin dollarin vuotuisen toistuvan liikevaihdon. Käyttö kolminkertaistui viidessä kuukaudessa ja kasvua veti hiekkalaatikoiden kysyntä. Launch HN -esittelyssä Agent37 kertoo asiakkaidensa maksavan keskimäärin 0,37 dollaria kuukaudessa hiekkalaatikkaa kohden. Molempien yhtiöiden luvut ovat niiden omia. Python 3.15 toi PEP 810:n mukaiset eksplisiittiset laiskat tuonnit, joissa esimerkiksi `lazy import json` tuo moduulin vasta ensimmäisellä käyttökerralla. Raskaita kirjastoja käyttävien ML-työkalujen käynnistysaika kuluu usein juuri tuonteihin. Python 3.10:n tuki päättyi samalla.

08 Benchmarkit ja vertailut

Artificial Analysisin Intelligence Indexiin tuli viikon aikana seitsemän uutta merkintää. Kärjessä ovat Metan Muse Wonder Full (Max, 56 pistettä) ja Muse Wonder (Max, 52). Muut merkinnät ovat Muse Spark Open (xhigh, 39), Reflectionin Beam 1.0 (30), Upstagen Solar Pro 4 (27), NVIDIAN Nemotron 3.5 Super VL (26) ja Nemotron 3.5 Bolt 74B-A9B (21). Päättelyasetukset vaihtelevat merkinnästä toiseen, joten pisteet eivät ole suoraan vertailukelpoisia. Boltia kannattaa verrata malleihin, joiden laskentakustannus on samaa luokkaa, eli noin yhdeksän miljardin aktiivisen parametrin malleihin, eikä listan kärkeen.

Epoch AI julkaisi kaksi mittausta, jotka kertovat kehityksen nopeudesta eri tavoin. Cyber ECI -indeksissä avointen mallien kärjessä on Z.ai:n GLM-5.3, jonka tulos vastaa suljettujen mallien tasoa kuuden tai seitsemän kuukauden takaa. Indeksillä on kybertehtäviin rajattu versio Epochin ECI-kyvykkyyksindeksistä, joka yhdistää useita vertailuja yhdelle asteikolle. Uusi SpireClimb puolestaan mittaa, miten mallit pelaavat Slay the Spireä ja sen jatko-osaa. Korttipakkaa rakentava peli vaatii pitkän aikavälin suunnittelua satunnaisuuden keskellä. Vielä muutama kuukausi sitten kärkeimallit pääsivät pelin hädin tuskin läpi. Nyt ne voittavat helpoimmalla tasolla yli 90 % ja vaikeimmalla jopa 20 % peleistä.

Kaksi viikon kirjoitusta kysyi, mitä indeksit ylipäättään mittaavat. The Batchin mukaan mallit suoriutuvat tehtävistä, joiden tuloksen voi tarkistaa, selvästi paremmin kuin harkintaa vaativista tehtävistä ja ero on suurempi kuin yleensä oletetaan. Vertailuindeksit koostuvat pääosin tarkistettavista tehtävistä, eli juuri niistä, joissa mallit ovat vahvimmillaan. Mercorin Labor Index -tutkimuksen benchmark-kartta oli lauantaina Hacker Newsin kommentoiduin tekoälyaihe (577 pistettä, 524 kommenttia). Sen mukaan 90 % yhdysvaltalaisista työtunneista kuluu tehtäviin, joita nykyiset benchmarkit eivät kata hyvin. Luku riippuu siitä, miten tehtävät luokitellaan ja mitä hyvä kattavuus tarkoittaa. Mercorin liiketoiminta perustuu asiantuntijoiden tuottamaan koulutus- ja arviointidataan, jota se myy laboratorioille.

Mittaaminen on siirtymässä yksittäisistä malleista kokonaisiin järjestelmiin. LMArenan Agent Arenan kärkeen nousi Max, reititin, joka valitsee kullekin kehotteelle sopivan mallin. Listan kärjessä on siis järjestelmä eikä yksittäinen malli. MLCommons valittiin vertailutiimiksi Yhdysvaltain National Science Foundationin tukemaan tehokkuushaasteeseen, jossa rakennetaan MLPerf Agents. Kahden vuoden ja 3,5 miljoonan dollarin rahoituksella syntyvä vertailu mittaa kokonaisten agenttityönkulkujen tehokkuutta yksittäisten mallikutsujen sijaan. Rajapintapalveluja mittaava MLPerf Endpoints v1.0

ottaa tuloksia vastaan jatkuvasti kiinteiden kierrosten sijaan ja tehonkulutus on raportoitava. MLPerf on mukana myös EU:n tekoälyjättitehtaiden hankintakilpailussa. MLCommons julkaisi lisäksi ostajan oppaan, joka kytkee viisi mittaria suoraan hankintapäätöksiin.

Uusista avoimista vertailuista Parsimony mittaa, kuinka paljon koodia mallit lisäävät. Nykyiset vertailut mittaavat ratkaisuasutetta sekä tokenien määrää ja hintaa, mutta ylimääräisestä koodista ne eivät rankaise. Parsimony laskee DeepSWE:n, SWE-bench Verifiedin ja SWE-bench Liven jo julkaistuista patcheista syntaksielementtejä, kuten lauseita, kutsuja, nimiä ja literaaleja. Rivimäärää ei käytetä, jotta koodin ahtaaminen harvemmille riveille ei toisi etua. Poistettu koodi vähennetään, joten tulos kuvaa nettomäärää. GPT-6:lle ja Claude 5.5:lle ei ollut julkisia patcheja, joten tekijä ajoi ne itse. Niiden ajoympäristö poikkeaa siksi muiden mallien julkaistuista ajoista.

Osa viikon kokeista mittaa yhtä ympäristöä, eivätkä niiden tulokset yleisty muihin tehtäviin. Opperin Jevmanissa kuusi

päätösmallia pelasi kukin 100 peliä Pac-Mania bottihaamuja vastaan. Compo AI:n DishwasherBench kysyy, osaavatko kärkeimmallit täyttää astianpesukoneen. Microsoft Researchin StudentSim simuloi yksittäisiä oppijoita, joiden palautteella tekoälytutoreita arvioidaan ennen kuin niitä kokeillaan oikeiden oppijoiden kanssa. Tulokset ovat juuri niin hyviä kuin simulaatio on. Simulaatiota vastaan optimoitu tutor oppii ennen kaikkea palvelemaan simulaatiota.

Viikon vilkkaimpiin keskusteluihin kuului arXiv-artikkeli 2610.13928 AI models can provably hide arbitrary capabilities (265 pistettä, 131 kommenttia). Kyvyn puuttumista ei voi osoittaa mustan laatikon testauksella, jos sen voi todistettavasti kätkeä. Väitteen laajuus ja oletukset selviävät vasta artikkelista. METR julkaisi samalla viikolla viisi kirjoitusta ulkopuolisesta arvioinnista. Kolmessa niistä ehdotetaan, miten arvioija voisi tutkia julkaisemattomien mallien linjausta, yhtiöiden sisäisten agenttien kontrollia ja valvontaa sekä näyttöä siitä, että tutkimus- ja kehitystyön automatisointi nopeuttaa kehitystä.

09 Työskentelytavat ja agentit

Anthropicissa yhdistettiin toukokuussa noin 2 500 muutosta päivässä ja syyskuussa yli 30 000. Yhtiö kertoo rakentaneensa oman vaihtoehdon Gitille, koska yhdistettyjen muutosten määrä kasvoi neljässä kuukaudessa 12-kertaiseksi. Pullonkaula ei ole enää koodin kirjoittaminen vaan versionhallinta ja sen jälkeen katselmointi, kun muutoksia yhdistetään kymmeniätuhansia päivässä.

Claude Code -tiimin kuvaus omasta katselmointikäytännöstään keräsi Hacker Newsissa 445 pistettä ja 436 kommenttia. Katselmointi on siirtynyt PR-vaiheen portista jatkuvaksi prosessiksi. Useimmissa muutoksissa Claude on katselmoinut ja korjannut koodin ainakin kerran ennen kuin ihmistä pyydetään katsomaan sitä. Tarkistukset kohdennetaan riskin mukaan. Jokaiselle muutokselle valitaan automaattinen tai ihmisen tekemä tarkistus sen mukaan, kumpi todennäköisimmin löytää juuri sen muutoksen tyypillisen vian. Lainatussa osuudessa ei ole mittareita. Vastakkainen esimerkki on The Registerin uutinen PostgreSQL 19:n viivästymisestä. Koneelliset katselmoijat löysivät enemmän vikoja kuin ihmiset ehtivät korjata. PostgreSQL:ssä muutokset hyväksyvät comitterit, eikä heidän määränsä kasva koneellisten löydösten tahdissa. Anthropicin AI Engineer New Yorkissa esittelemät dynaamiset työnkulut ottavat mallia Unixista. Iso työ kootaan pienistä ja rajatuista osista, joiden tuotokset voi tarkistaa. Latent Spacen raportin mukaan pullonkaula on kehittäjän kyky katselmoida, ei agenttien määrä.

Toinen viikon havainto koskee sitä, missä agentin ohjaus tapahtuu. Ohjelmistokehittäjän kirjoitus nobody should type code by hand anymore. here's what's left. (2026) kuvaa tilanteen, jossa agentti ohittaa diagnostiikkakirjaston. Kirjoittaja ei korjaa AGENTS.md-tiedostoa, koska ohjeen ja agentin käytöksen välille syntyy ajan mittaan eroa. Sen sijaan pre-commit-koukku tarkistaa git diffistä suorat observability-kutsut, hylkää commitin ja antaa agentille kehoitteen käyttää kirjastoa. Hylkäysviesti toimii samalla kehoitteena, joka kertoo korjauksen juuri siinä kohdassa, jossa virhe syntyi. Pelikehittäjän kirjoitus agenttiympäristön eli harnessin suunnittelusta Claude Coden kanssa (172 pistettä) vie painopisteen samaan suuntaan: työkaluihin, testeihin ja palautesilmukoihin, joiden kautta agentti näkee työnsä tuloksen. NVIDIA kuvasi kaksi agenttiputkea, joissa varmistettavuus on keskeistä. Ensimmäisessä agentit kirjoittavat Apache Sparkin UDF-funktioita uudelleen GPU-kiihdytettäväksi. Toisessa piirisuunnittelun agentteja parannetaan NeMo Gym -ympäristöllä. NVIDIAN mukaan agentit tuottavat uskottavan näköisiä mutta virheellisiä tuotoksia.

Datasette-projektin kehittäjä rakensi viikon alussa puhelimelta Codex Remoten kautta web-edit-sovelluksen. Sillä hän pystyi muokkaamaan tunnustiedostoa ilman, että tunnukset kulkevat kielimallin läpi. Sovellus toimii Tailscale-verkon yli. Agentti voi käynnistää sen ja sen jälkeen ihminen muokkaa tiedostoja suoraan CodeMirror-editorissa. Sovelluksen seuraavassa

versiossa työ jaettiin kahden valmistajan mallien kesken. GPT-6 Astra Extra High ajoi Codexissa Python 3.15:n uuden profiloijan sovellusta vastaan ja kirjoitti profilointiraportin. Raportin pohjalta Claude Opus 5.5 teki suorituskykyoptimoinnit versioon 0.1a1. Raportti toimi mallien välisenä luovutusdokumenttina, jonka ihminen voi lukea ennen kuin toinen agentti jatkaa sen pohjalta. Saman kehittäjän Tree Appreciator -peli syntyi Clauden iPhone-sovelluksen tavallisessa keskustelutilassa ilman erillistä koodausagenttia. Opus 5.5 Max työsti three.js-peliä kolme tuntia ja seitsemän minuuttia. Kehote pyysi mallia miettimään tarkkaan, miten pelistä saa sekä hauskan että kauniin. Kirjoittaja toteaa, että havainto pyynnön toimivuudesta on hieman nolo hänen tuoreen puheensa valossa.

Kolmas kehityslinja näkyy työkalujen julkaisuissa: agenteista on tulossa palveluita, joita ajetaan ilman valvontaa. Claude Code 2.1.301 lisäsi valvomattomaan uudelleenyritystilaan muuttujan

`CLAUDE_CODE_RETRY_WATCHDOG_MAX_USAGE_LIMIT_WAIT_MS`.

Sen avulla pyyntö lopetetaan, jos käyttöraja nollautuu vasta liian kaukana tulevaisuudessa. Cline 4.1.25:n Cloud sessions ajaa tehtävän isännöidyssä hiekkalaatikossa valitusta GitHub-reposta ja -haarasta. Tehtävä jatkuu, vaikka editori suljettaisiin. Kaikki Cline:n perjantain julkaisut yrittävät palveluntarjoajan kesken streamin tulleet virheitä uudelleen enintään kolme kertaa ennen kuin vuoro hylätään. Codex CLI 0.163.0 sallii jatkaviestin lähettämisen kesken vastauksen. Keskeytys pysäyttää myös aktiiviset Code Mode -solut ja niiden sisäkkäiset työkalukutsut. Uusi `/daemon`-komento käynnistää hallitun taustapalvelimen uudelleen, mutta varoittaa ensin työstä, jonka uudelleenkäynnistys voi keskeyttää. Gemini CLI korjasi viikon aikana jatkettujen istuntojen vikoja. Työkaluvastauksia ei enää kahdennu, historia ei katoa nopeassa poistumisessa eivätkä nopeat Ctrl-C-keskeytykset enää riko istuntoa.

Claude Code 2.1.302 korjasi Prompt is too long -virheen, joka syntyi, kun vuoro pysäytettiin kesken kontekstin tiivistyksen. Sama versio toi muuttujan `CLAUDE_CODE_AUTO_MODE_MODEL`, jolla valitaan asiakaspuolen auto-tilan luokittelijan malli. Luokittelija päättää, mitkä toimenpiteet ajetaan kysymättä käyttäjältä. Mallin valinnassa punnitaan viivettä, kustannusta ja tarkkuutta. Cline:n oletusmallit vaihtuivat viikon aikana nopeammin kuin moni käyttäjä ehtii huomata. Tiistain Cline 4.1.24 käytti Cline- ja OpenRouter-palveluntarjoajilla oletuksena Claude Sonnet 5.5:tä, kun mallia ei ollut valittu. Perjantain 4.1.25 vaihtoi Cline-palveluntarjoajan oletuksen

Gemini 4 Argoniin. Lauantaina Sonnet 5.5 palasi oletukseksi ja suositusten kärkeen. Käytön mukaan maksavan kannattaa varmistaa, että oma mallivalinta on tallennettu, koska oletuksen vaihtuminen muuttaa kustannuksia huomaamatta. Cursor julkaisi Origin App Marketplacen ja toi pilviagenteille canvas-näkymät.

Ilman valvontaa toimiva agentti on jäljitettävä ja OpenTelemetry-tuki lisääntyi viikon julkaisuissa. Claude Platform sai beetana OpenTelemetry-viennin, jolla Messages API -pyynnöt ja Claude Managed Agents -istunnot voi jäljittää omaan havaittavuusjärjestelmään. SDK tuottaa spaneja seitsemällä ohjelmointikielellä. Claude Coden `api_request` -tapahtuma erittelee välimuistikirjoitukset kenttiin `cache_creation_5m_tokens` ja `cache_creation_1h_tokens`. Erittely helpottaa kustannusten seurantaa, koska tunnin välimuistikirjoitus on hinnoiteltu viiden minuutin kirjoitusta kalliimmaksi. Cline:n hub-daemon julkaisee istunnon käynnistäneen asiakkaan ja yhteystavan OTel-resurssiattribuutteina. Jaettua hubia käyttävä tiimi näkee siten, mistä asiakkaasta ja minkä yhteyden kautta kukin istunto käynnistettiin. Gemini CLI:n ACP-silta välittää nyt käyttötiedot `usage_update`-ilmoituksina.

Tekoälyavusteinen työ näkyy jo infraprojektien commit-tiedoissa. Llama.cpp:n viikon committeissa on merkintöjä kuten Assisted-by: OpenCode, Assisted-by: Codex ja AI-usage: full (Claude + Sonnet 5.5). Merkintä AI-usage: draft erottaa tekoälyn luonnostelemat muutokset kokonaan sen kirjoittamista. CUDA-ympäristömuuttujien siivoukseen on merkitty avustajaksi `pi:llama.cpp/Qwen3.8-27B`. Llama.cpp:llä ajettu paikallinen malli on siis osallistunut llama.cpp:n omaan kehitykseen. Lauantain web-käyttöliittymämuutoksessa on merkintä Assisted-by: pi:zai-org/GLM-5.3. vLLM:n julkaisukandidaateissa on rivejä Co-authored-by: Claude Opus 5.5 ja Co-authored-by: Codex. Phoronixin mukaan myös Linux-ytimen luoja työstää Claude Codella mittavaa parannusta ytimeen. Ytimen kehittäjäyhteisössä tekoälyavusteiset kontribuutiot ovat olleet kiistanalaisia.

Pitkien agenttijaon mittakaavasta kertoi ensimmäisen persoonan räiskintäpelin dekompilointi, johon kului 500 miljardia tokenia (205 pistettä, 147 kommenttia). Tällaisessa ajossa kustannus riippuu ratkaisevasti siitä, kuinka suuri osa syötteestä osuu välimuistiin. Cockroach Labs raportoi viiden kuukauden kokeilusta, jossa vikoja käsiteltiin potilaina ja koodausagentteja lääkäriyhtymänä. Kokeilun pituus erottaa sen tavallisista viikonlopun demoista.

10 Liitännäiset, MCP ja repot

MCP-määrittelyn viikon tärkein muutos oli dokumentaatiokorjaus, jolla on käytännön vaikutuksia (#3427). Version 2026-07-28 pikaohje Build an MCP server kytki TypeScript-palvelimen kutsulla `server.connect(new StdioServerTransport())`. Kutsu palvelee vain asiakkaita, jotka avaavat yhteyden 2025-aikakauden `initialize`-kättelyllä. Ohjeen mukaan rakennettu palvelin epäonnistui versioneuvoittelussa 2026-07-28-määrittelyä käyttävien asiakkaiden kanssa. Virhe jäi huomaamatta, koska Inspectorin oletusaikakausi on `legacy`. Korjattu ohje rakentaa palvelimen `createServer()`-tehtaalla ja käynnistää sen SDK:n kutsulla `serveStdio(createServer)`. Vian tunnistaa siitä, että palvelin läpäisee Inspectorin oletustestit mutta kaatuu versioneuvoitteluun uudemman asiakkaan kanssa. Testaus kannattaa siksi tehdä Inspectorilla, jonka aikakaudeksi on asetettu uusi määrittely. Yhteisössä Primitive Grouping -kiinnostusryhmä muutettiin Progressive Disclosure -työryhmäksi ja sen peruskirja hyväksyttiin. Nimi viittaa periaatteeseen, jossa mallille näytetään aluksi vain tiivis kuvaus työkaluista ja yksityiskohdat ladataan tarvittaessa. Viikon muut commitit olivat riippuvuuspäivityksiä ja linkkikorjauksia.

MCP-tuki levisi viikolla suunnittelu- ja luoviin työkaluihin. SolidCraft (123 pistettä) on Rustilla kirjoitettu avoimen lähdekoodin parametrinen CAD-ohjelma, jonka piirteitä ja mittoja agentti voi muokata samaan tapaan kuin koodia. Hanke on kokeilun arvoinen niille, jotka haluavat agentin tuottavan mitoitettuja osia, vaikka tuotantovalmiudesta ei ole näyttöä. Kokeellinen äänityöasema Blokkily on toistaiseksi konsepti. Itse ylläpidettävä Edi Life OS on suppea harrastusprojekti, mutta sen uhkamallista voi oppia. Ilman asiakas kohtaista valtuutusta MCP-palvelin avaa henkilökohtaisen datan jokaiselle siihen kytketylle asiakkaalle.

Skillit vakiintuivat viikolla muodoksi, joka toimii eri työkaluissa. Googlen android/skills on Android-kehityksen virallinen kokoelma agent skills -standardin mukaisia SKILL.md-tiedostoja, jotka perustuvat developer.android.comin ohjeisiin. Repo keskittyy tilanteisiin, joissa arvioinnit osoittavat kielimallien suoriutuvan heikosti. Asioita, jotka mallit jo osaavat, se ei priorisoi. Arviointeihin perustuva rajausta tekee siitä kokeilun arvoisen kaikille, jotka tekevät Android-kehitystä agenttien kanssa. K-Dense AI nimesi Claude Scientific Skills -kokoelmansa uudelleen scientific-agent-skillsiksi. Uusi nimi kertoo, että kokoelman 177 validoitua taitoa ja yli 100 tieteellistä tietokantaa toimivat avoimen Agent Skills -standardin kautta Cursorissa, Claude Codessa, Codexissa, Pi:ssä ja Antigravityssä. Taustalla on arXiv-julkaisu 2609.00065. Tekijöiden väitettä yli 250 000

käyttäjistä ei ole tarkistettu. Elämäntieteissä työskenteleville repo on kokeilun arvoinen.

Claude Coden mod-rajapinta laajeni versiossa 2.1.300. Uusi `virtualText` piirtää kehoteruutuun tekstiä, joka ei koskaan päädy luonnokseen eikä lähetettyyn kehoitteeseen. Kutsu `$.ui.view({ agentId })` näyttää valitun agentin keskusteluhistorian päänäkyssä. Sunnuntaina awesome-claude-code-listalle perustettiin oma Mods-luokka, jonka ensimmäinen merkintä on prismantis. Listalle viikolla lisätyt PLUR, funes, pstack-claude ja Mycelium keskittyvät istuntojen yli säilyvään muistiin ja usean agentin koordinointiin. Listamerkinnän perusteella niiden laatua ei voi arvioida.

Viikon nopeasti nousseista repoista muutama ansaitsee arvion. Paikallinen hakuindeksi xerj indeksoi kansion koodin, dokumentit, lokit ja PDF:t yhdellä komennolla ja tarjoaa Elasticsearch-yhteensopivan rajapinnan. Projektin oman arvion mukaan agentti kuluttaa haussa 40 kertaa vähemmän tokeneita kuin lukiessaan tiedostoja yksitellen grepillä ja sedillä. Työkalu on kokeilun arvoinen suurissa monorepoissa, joissa tiedostojen lukeminen on todellinen kustannuserä. Indeksiiin päätyvät myös lokit, joten sen sijainti ja käyttöoikeudet kannattaa tarkistaa ennen kuin agentti pääsee käyttämään sitä. Sunnuntaina trendilistalle nousi stablyai/orca. Se ajaa Codexia, Claude Codea, OpenCodea ja Pi:tä rinnakkain, kukin omassa worktreessään ja kokoa ne yhteen näkymään. Puhelinsovellus ilmoittaa agentin valmistumisesta ja välittää jatkoviestit. Orca sopii kokeiluun usean agentin rinnakkaiseen työhön omilla tilauksilla, vaikka sen lupaus "100x builders" on markkinointia. Samaan käyttöön liittyy myös terminaalikehittäjän ehdotus OSC 7501 -ohjaussekvenssistä, jolla ohjelma voi ilmoittaa tilansa terminaalille (186 pistettä). Ehdotus muuttuu käytännöksi vasta, jos terminaalit ja työkalut ottavat sen käyttöön.

Cloudflaren ilmaisostasolla toimiva itse ylläpidettävä agentti talorys (325 pistettä) on kokeilun arvoinen, kunhan ilmaistason suoritusrajat ja kielimallin kustannukset tarkistetaan ja agentin tunnukset rajataan suppeiksi. Alishahryar1/free-claude-code reitittää Claude Coden, Codexin ja usean muun ympäristön pyynnöt 61 palveluntarjoajan kautta ja lupaa yli 1,3 miljardia ilmaista tokenia kuukaudessa. Hanke ilmoittaa itse, ettei Anthropic ole hyväksynyt sitä. Sitä voi kokeilla koodilla, joka ei ole luottamuksellista, mutta asiakasprojekteihin se ei sovi. iFixAi lupaa auditoida agentin alle 120 sekunnissa, mutta README on markkinointipainotteinen eikä työkalua ole validoitu riippumattomasti. Sitä voi kokeilla testiagentilla, mutta tuotannon tarkastukseen se ei vielä sovellu.

Pitkäkestoisia Claude Code -silmukoita hallitseva Spool on vielä kokeiluvaiheessa. Xcoden korvaava xtool ei ole tekoälyprojekti. Se on silti hyödyllinen iOS-kehittäjälle, joka haluaa rakentaa sovelluksia esimerkiksi CI-putkessa tai agentin Linux-hiekkalaatikossa. Paikallinen PDF-muunnin Sygal on kokeilun arvoinen RAG-esikäsitteilyssä, kunhan sen tarkkuus todennetaan omalla aineistolla. Laitteisto oli toistuva teema

yhteisön nostoissa. GNOME-blogin kirjoitus tekoälystä laitteiston takaisinmallinnuksessa sai 250 pistettä. Konnect tuo tekoälyavusteisen piirilevy-suunnittelun KiCad 10:een. Llama.cpp:llä jailbreakatulla PS5:llä GGUF-malleja ajava PS5LM kertoo lähinnä llama.cpp:n siirrettävyydestä, eikä se ole käytännöllinen ajoalusta.

11 Tietoturva ja haavoittuvuudet

Viikon opettavaisin tietoturvavika ei ollut haavoittuvuus vaan tarkastus, joka ohitti koodia ilmoittamatta siitä. Anthropicin virallisen liitännäiskokeelman security-guidance-liitännäisestä korjattiin 12.–14.10. joukko vikoja, joiden vuoksi tarkastus saattoi raportoida puhtaan tuloksen koodista, jota se ei ollut nähnyt. Diff-jäsenen pilkkoi syötteen jokaisesta merkkijonon `diff --git` esiintymästä. Tiedoston sisällössä esiintynyt merkkijono katkaisi tiedoston siihen kohtaan, eikä loppuosaa tarkastettu lainkaan. Jäsenen pudotti lisäksi tiedostot, joiden nimet git ympäröi lainausmerkeillä. Näin tapahtuu, jos nimessä on lainausmerkki, kenoviiva tai ohjausmerkki. Korjaus (#6542, yhdistetty 14.10.) jakaa syötteen vain otsikkorivien kohdalta ja purkaa lainausmerkityt nimet. Stop-hook raportoi jäsentämättömät otsikot laskurina `diff_hdr_unparsed`, joka näytetään vain, kun sen arvo on muu kuin nolla. PR #6375 varmistaa, ettei epäonnistunutta tarkastusta raportoida puhtaana.

Toinen vika liittyi siihen, miten löydökset välitettiin. Commit- ja push-löydökset menivät vain `hookSpecificOutput.additionalContext`-kenttään, jota Claude Coden `asyncRewake`-polku ei lue. Polku näyttää Claudelle `stderr`in, tai `stdout`in, jos `stderr` on tyhjä. Siksi Claude sai löydösten sijaan varoituksen aina, kun sisäinen CLI kirjoitti varoituksen `stderr`iin (`claude-code#99839` ja `#92987`). PR #6379 ohjaa löydökset kaikissa tapahtumissa `stderr`iin ja sisäisten työkalujen lokirivit liitännäisen debug-lokiin. PR #6378 lähettää `ANTHROPIC_CUSTOM_HEADERS`-otsakkeet myös liitännäisen omissa API-kutsuissa, mikä koskee yhdyskäytävän kautta reititettyjä asennuksia. Uusi telemetria-avain `gateway_cfg` kertoo lukuvarvolla, osoittaako `ANTHROPIC_BASE_URL` muuhun isäntään kuin `api.anthropic.com` ja lähetetäänkö mukautettuja otsakkeita. Isäntää tai otsakkeiden nimiä avain ei sisällä. Korjauksen saa päivittämällä liitännäisen kokoelmasta. Lähteet eivät mainitse versionumeroa. Ennen 13.–14.10. saadut puhtaat tulokset kannattaa ajaa uudelleen, erityisesti yhdyskäytävän takana toimivissa asennuksissa sekä commit- ja push-tarkastuksissa.

Cline muutti liitännäisten eristystä kahdesti saman viikon aikana. Tiistain CLI 3.0.71 ja SDK 0.0.93 siirsivät liitännäiset oletuksena hubin omaan prosessiin yhden rekisterin alle.

Aiempi aliprosessimalli jäi käyttöön vain asetuksella `CLINE_PLUGIN_MODE=sandbox`. Rekisteri suojaa liitännäisen importin, setupin, koudut, työkalut ja komennot aikakatkaisuilla ja seuraa kunkin liitännäisen tilaa. Epäonnistunut liitännäinen näkyy siksi `plugin_failure`-merkintänä, eivätkä sen työkalut katoa huomaamatta. Gemini CLI:n saman päivän nightly jatkaa muiden laajennusten lataamista, vaikka yksi kaatuisi sallittujen listan tarkistuksessa (#29720). Perjantain SDK 0.0.94 ja CLI 3.0.72 poistivat `in_process`-tilan kokonaan, samoin sen hiljaisen varakäytöksen. Oletuksena ovat nyt worker-säikeet (`CLINE_PLUGIN_MODE=worker`) ja vaihtoehtona on aliprosessihiekkalaatikko (`sandbox`). Jokainen liitännäisen lataus kaatuu virheeseen `PluginBootstrapMissingError`, jos isäntä ei toimita `plugin-sandbox-bootstrap.js`-tiedostoa. Virhe nimeää muuttujan `CLINE_PLUGIN_BOOTSTRAP_PATH`. SSH-etähub ei vielä toimita bootstrap-tiedostoa, joten sen kautta liitännäisten työkalut näkyvät puuttuvina. Ympäristöön asetettu arvo `in_process` ei enää toimi, joten se on vaihdettava arvoon `worker` tai `sandbox`. Node.js:n worker-säikeet jakavat saman prosessin, joten aidosti epäluotettavien liitännäisten eristykseen sopii `sandbox`.

Gemini CLI kiristi luottamusrajoja kolmessa kohdassa. Versio 0.64.0 korjasi sen, miten kansion luottamustila välitetty headless-tilaan (#29031, PR #29528). Esijulkaisu 0.65.0-preview.0 pakottaa työtilan asetukset vain luku-tilaan epäluotetuissa kansioissa (#29583). Se myös tarkistaa OAuth-paluukutsun `iss`-parametrin RFC 9207:n mukaisesti (#29616). RFC 9207 määrittelee, miten valtuutuspalvelin ilmoittaa tunnisteensa. Tarkistus torjuu mix-up-hyökkäyksiä, joissa asiakas sekoittaa valtuutuspalvelimet keskenään. Epäluotettuja repoja CI:ssä tai headless-tilassa käsittelevien kannattaa päivittää vähintään versioon 0.64.0. Loppuviikon nightly tarkistaa lisäksi päivityksen jälkeen, että käytössä oleva binääri todella vaihtui (#29757). Aiemmin `PATH`:ssa ensimmäisenä oleva toinen binääri saattoi peittää päivitetyn version. Päivitys raportoitui silloin onnistuneeksi, vaikka käyttäjä ajoi edelleen vanhaa versiota.

n8n 2.44.0 ulotti solmutyypikäytännöt myös agenttisolmun työkaluihin (#39711). Korjauksen otsikon perusteella agentin työkalukutsut ovat aiemmin jääneet näiden käytäntöjen ulkopuolelle. Tunnustietotyyppien käytäntö koskee nyt myös tunnusten testejä ja luotauksia (#39927). Versiot 2.43.6 ja 2.44.2 luovat webhookin pyyntörungolle oman XML-jäsentimen jokaista pyyntöä varten (#41113 ja #41112). Julkaisutiedot luokittelevat muutoksen virhekorjaukseksi, eikä sille ole CVE-tunnistetta. Rinnakkaisten pyyntöjen jaettu jäsentimen tila voi kuitenkin sotkea pyyntöjen käsittelyä, joten XML-webhookeja käyttävien kannattaa päivittää.

Llama.cpp:ssä korjattiin viikon aikana useita muistin käsittelyyn liittyviä virheitä, joita ei merkitty tietoturvakorjauksiksi. Metal-taustan `mul_mat_id`-polku oletti, että tokenilla on enintään yksi osuma ja jokaisella asiantuntijalla kiinteä kapasiteetti. Samojen id:iden toistuessa tulosrivejä saattoi tulla vääriä tai puuttua ja kirjoitus saattoi ylivuotaa seuraavan asiantuntijan viipaleeseen. Korjaus (b11569) rakentaa jokaiselle asiantuntijalle tiiviin listan kahdella läpikäynnillä. Apple Siliconilla MoE-malleja ajavien kannattaa varmistaa, että käytössä on vähintään build b11569, koska virhe tuottaa vääriä tuloksia huomaamatta. SYCL-taustassa rivien loppupäiden DMMV-lukujen rajat tarkistetaan nyt (b11565) ja `mul_mat`-, `split buffer`- ja `host pool`-koodin muistivirheet korjattiin (b11693). Lauantain b11714 lisäsi web-käyttöliittymän MarkdownContent-komponenttiin valinnaisen `allowHtml`-ominaisuuden. Ominaisuuden ollessa päällä Markdownin raaka-HTML-renderöidään `DOMPurify`-sanitoinnin jälkeen sen sijaan, että se näytettäisiin tekstinä. Oletuksena käytös on ennallaan. Mallin tuottama HTML on XSS-hyökkäyspintaa ja sanitoinnin ohitukset ovat tunnettu virheluokka. Siksi ominaisuus kannattaa pitää pois päältä jaetuissa palveluissa ja silloin, kun käsitellään epäluotettavaa sisältöä. Binaarijulkaisuihin liitetään GitHubin attestatit, joilla ladatun binaarin alkuperän voi todentaa.

Ollama 0.40.6-rc1 muutti rekisterin uudelleenohjausten DNS-käsittelyä. Välityspalvelimen takana uudelleenohjauskohteen nimi selvitetään välityspalvelimella. Ilman välityspalvelinta nimi selvitetään ja kiinnitetään paikallisesti kuten ennenkin. Oletusrekisterin ja CDN:n isäntänimiin luotetaan nyt ilman paikallista DNS-hakua. Open WebUI 0.12.3:n salatut keskustelut perustuvat selaimessa ajettavaan MLS-protokollaan. Palvelin välittää enintään 16 KiB:n viestit tallentamatta niitä ja osallistujat voivat varmentaa toisensa vertaamalla koodeja. Keskustelu päättyy, kun kanavan jäsenyys tai oikeudet muuttuvat. Rajoitus on sama kuin muussakin selainsalauksessa. Salaukoodin toimittaa sama

palvelin, jonka ei pitäisi nähdä viestejä. Suojaus kattaa tallennuksen ja palvelinlokitt, mutta ei ylläpitäjää, joka voisi vaihtaa selaimelle toimitettavan koodin. Anthropicin Compliance API:n uusi `user_preferences`-pääte piste tuo käyttäjien vapaamuotoiset pysyvät ohjeet compliance-viennin piiriin. Ohjeissa voi olla henkilötietoja. Organisaatioiden kannattaa siksi päivittää säilytys- ja käyttöoikeuskäytäntönsä ja rajata, mitkä avaimet saavat `read:compliance_user_data`-oikeuden.

Viikon agenttitapauksissa ei ollut kyse haavoittuvuuksista, joilla olisi CVE-tunniste. Tivin raportoimassa tapauksessa agentti jakoi toimitusjohtajan tilitiedot työntekijöille. Toimitusjohtajan mukaan agentit voivat toimia kaoottisesti ilman hakkerointia tai tahallista väärinkäyttöä. Tällainen tilanne syntyy tyypillisesti, kun agentti toimii yhden henkilön tunnuksilla mutta palvelee monta kysyjää. Ruotsalaisen TV4:n ohjelma siirtyi vuoteen 2027, kun tekoälyagentti hävitti editoitua materiaalia. CSO Onlineen jutun mukaan agentit eivät ole karkaamassa käsistä, vaan karkkilaboratoriot laiminlyövät tietoturvan perusasiat. Juttu keräsi Hacker Newsissa 933 pistettä ja 603 kommenttia. Kontrollit ovat tuttuja: agentille omat ja suppeat tunnukset, kysyjän henkilöllisyyteen sidottu valtuutus, ulospäin suuntautuvan verkkoliikenteen raja, työkalukutsujen lokitus ja ihmisen hyväksyntä ennen peruuttamattomia toimenpiteitä.

Perinteisen tietoturvan puolella Kyberturvallisuuskeskus raportoi kaksi verkon reunalaitteiden haavoittuvuutta, joita hyödynnetään aktiivisesti. Ivanti Connect Securen haavoittuvuuteen CVE-2026-104935 on julkaistu päivitys, jonka versionumero on tarkistettava Ivantin tiedotteesta. Pelkkä päivitys ei poista takaportteja, jotka hyökkääjiä on mahdollisesti jo asentanut. Siksi tarkistukseen kannattaa käyttää valmistajan eheystyökalua ja käydä lokit läpi. Cisco Secure Email Gatewayn kriittinen haavoittuvuus (tiedote 2026-36) mahdollistaa komentojen suorittamisen laitteen käyttöjärjestelmässä pääkäyttäjän oikeuksin. Tiedotteen tiivistelmä ei kerro CVE-tunnistetta eikä korjattua versiota. Itse ylläpidetyt mallipalvelut julkaistaan usein juuri VPN-yhdyskäytävien ja palomuurien takaa, joten reunalaitteiden päivitysviive koskee myös niitä. Epoch AI:n arvio, jonka mukaan avoimet mallit ovat kybertaidoissa kuusi tai seitsemän kuukautta suljettujen jäljessä, on uhkamallinnuksessa käyttökelpoinen nyrkkisääntö. Kevään 2026 suljettujen karkkimallien kybertaitojen kannattaa olettaa olevan nyt saatavilla myös paikallisesti ajettavissa malleissa ilman palveluntarjoajan käyttörajoituksia.

Lähteet

Gemini 4 Argon: Built to solve your hardest problems · Google DeepMind

Gemini 4 Argon for Robotics · Google DeepMind

Developing Gemini 4 Argon responsibly and safely · Google DeepMind

gemini-3.8-flash · Gemini API changelog

[AINews] AI Engineer NYC: Gemini 4 Argon, Vals AI Benchmarks · Latent Space

llm-gemini 0.35 · Simon Willison

AINews: Meta Muse Wonder & Wonder Full, MSL's Claude 5.5-tier Frontier Model · Latent Space

Muse Wonder (Max): Intelligence Index 52 · Artificial Analysis

Muse Wonder Full (Max): Intelligence Index 56 · Artificial Analysis

llm-meta-ai 0.2 · Simon Willison

Welcome Muse Spark Open: open-weight frontier multimodal mixture-of-experts model from Meta · Hugging Face -blogi

Scale Agentic Coding Workflows with Meta's Muse Spark Open on NVIDIA GB300 NVL72 · NVIDIA Technical Blog

Muse Spark Open, the Most Capable US Open-Weight Model to Date · Hacker News: MCP

Meta shakes up the frontiers, open and closed · Interconnects

Muse Spark Open (Xhigh) Intelligence Index: 39 · Artificial Analysis

AI models may be capable of self-replicating without access to a datacenter · Anthropic (Hacker News)

AI models may soon self-replicate without access to data centers · Reuters (Hacker News)

Claude Platform release notes, October 15, 2026 · Anthropic API release notes

Claude Platform release notes, October 14, 2026 · Anthropic API release notes

How Third Moment Research runs 10,000 agents for 36+ hours · OpenAI

Stack Overflow for Agents triples signups with ChatGPT plugin · OpenAI

Bringing GPT-Rosalind to academic research labs · OpenAI

Show HN: Jevman: AI decision models play Pac-Man · Hacker News: MCP

Introducing Timestral · Mistral

Claude Haiku 5.5, Mistral Large 4, and Local Models for Mac and Windows · The Batch (deeplearning.ai)

Introducing The X Subscription · xAI news

Introducing Grok Voice 3.0 · xAI news

Accelerating the search for antimicrobial peptides with Google Antigravity · Google DeepMind

Grounding AI agents in biological reality: Building blocks for autonomous wet labs · Google Research

Welcome Beam: a 501B open model you can own, customize, and scale · Hugging Face -blogi

Reflection has released Beam, their first model launch · Artificial Analysis

Beam 1.0 Intelligence Index: 30 · Artificial Analysis

Solar Pro 4 Intelligence Index: 27 · Artificial Analysis

NVIDIA Nemotron 3.5 Super VL Powers Specialized Agents with Multimodal Reasoning · NVIDIA Technical Blog

NVIDIA Releases Nemotron 3.5 Super VL and Nemotron 3.5 Bolt 74B A9B · Artificial Analysis

Nemotron 3.5 Super VL Intelligence Index: 26 · Artificial Analysis

Nemotron 3.5 Bolt 74B-A9B Intelligence Index: 21 · Artificial Analysis

Apertus 1.5 Release Notes · Hugging Face -blogi

Run fast, private agents on-device with LFM2.5-24B-A2B · Hugging Face -blogi

Trending: LiquidAI/LFM2.5-24B-A2B · HF trending -mallit

openbmb/MiniCPM-V-4.7-35B-A3B · HF trending -mallit

Trending: openbmb/MiniCPM-V-4.7-1B · HF trending -mallit

microsoft/BitNet · GitHub trending

We turned EmbeddingGemma 2 into a speech-to-text model · Hacker News

Welcome FLUX.3 Image Klein in Diffusers · Hugging Face -blogi

black-forest-labs/flux-3-image-klein · HF trending -mallit

black-forest-labs/flux-3-image-klein-comfy · HF trending -mallit

Trending: deepseek-ai/DeepSeek-V4.1-Flash-1017 · HF trending -mallit

abenzers/Qwen-Image-2.1-Uncensored-Quantized · HF trending -mallit

nerkyor/Qwen3.8-27B-Coder390-EfficientThink-Opus5.5-GPT6Astra-Grok4.7-DSV4Pro-K3-SFT-RL00-MTP-DFlash2 · HF trending -mallit

Retrofitting language models to operate over bytes · Hacker News: MCP

Scaling Indic Pretraining: 2.8 Trillion Tokens and an Open Data Generation Pipeline · Hugging Face -blogi

v0.40.4 · Ollama-julkaisut

Ollama v0.40.4 · Ollama-julkaisut

v0.40.5 · Ollama-julkaisut

v0.40.5-rc1: New model... (#19030) · Ollama-julkaisut

v0.40.6 · Ollama-julkaisut

v0.40.6-rc1: transfer: support proxies on registry redirect targets · Ollama-julkaisut

v0.40.7 · Ollama-julkaisut

v0.32.0 · vLLM releases

vLLM v0.32.0rc0 · vLLM releases

v0.32.0rc1: [Dependency] Update Humming and remove NVFP4 scale workaround (#61556) · vLLM releases

v0.32.0rc3: Fix NaN output of the Mamba2 chunk scan kernel on sm90 · vLLM releases

Granite Switch 4.1: Chat, Decisions, and Speech from One Endpoint · Hugging Face -blogi

v0.7.0 · llama.cpp-julkaisut

b11565 · llama.cpp-julkaisut

b11569 · llama.cpp-julkaisut

b11571 · llama.cpp-julkaisut

b11572 · llama.cpp-julkaisut

b11610 · llama.cpp-julkaisut

b11614 · llama.cpp-julkaisut

b11618: HIP: use 64-row MMQ tiles for some quant types on RDNA3.5 (#28195) · llama.cpp-julkaisut

b11642 · llama.cpp-julkaisut

b11644: kv-cache: honor PARTIAL_ONLY in the plain KV cache (#30514) · llama.cpp-julkaisut

llama.cpp b11693 · llama.cpp-julkaisut

llama.cpp b11695 · llama.cpp-julkaisut

llama.cpp b11699 · llama.cpp-julkaisut

llama.cpp b11700 · llama.cpp-julkaisut

b11707 · llama.cpp-julkaisut

b11709 · llama.cpp-julkaisut

b11710 · llama.cpp-julkaisut

b11713 · llama.cpp-julkaisut

b11714 · llama.cpp-julkaisut

b11715 · llama.cpp-julkaisut

b11716 · llama.cpp-julkaisut

b11717 · llama.cpp-julkaisut

b11718 · llama.cpp-julkaisut

ReplaySSM: Rethinking State Storage for Faster Hybrid-Model Decoding · NVIDIA Technical Blog

lyogavin/airllm · GitHub trending -syöte (mshibanami/GitHubTrendingRSS)

Intel D270 GPU AI Accelerator: Up to 480 GB LPDDR5X · Hacker News

langgraph=1.2.15 · LangGraph releases

langgraph=1.2.16 · LangGraph releases

langgraph-checkpoint-postgres=3.2.0 · LangGraph releases

n8n 2.43.7 · n8n releases

v0.12.1 · Open WebUI releases

Open WebUI v0.12.3 · Open WebUI releases

NVIDIA Run:ai is Now Open Source as Part of NVIDIA DSX AI Factory Platform · NVIDIA Technical Blog

lightseekorg/tokenspeed · GitHub trending

[AINews] Modal's \$1B ARR, \$16B Series D makes it NYC's newest AI Infra Decacorn, 3x usage and 3x valuation in 5 months, 15x in 1 yr · Latent Space

Launch HN: Agent37 (YC F26), Low-cost agent sandboxes for running agents · Hacker News

Hello Python 3.15, goodbye 3.10, and now I can use 3.11 · Simon Willison

Open-weight models trail closed models in cyber capabilities by 6–7 months · Epoch AI

AI systems are improving rapidly at Slay the Spire and its sequel · Epoch AI

90% of US working hours are spent on tasks not covered well by AI benchmarks · Hacker News: agents

Introducing Max in Agent Arena · LMArena

MLCommons Selected as Benchmarking Team for U.S. National Science Foundation-Backed AI Efficiency Challenge to Build MLPerf Agents · MLCommons

Open, peer-reviewed measurement for today's AI buyers: MLPerf in the EU's AI Gigafactories tender, and Endpoints v1.0 opens for submissions · MLCommons

What is MLPerf Endpoints, and why it matters to buyers · MLCommons

Show HN: Parsimony, an open-source benchmark for how much code AI models add · Hacker News: MCP

DishwasherBench: Can Frontier Models Stack a Dishwasher? · Hacker News

Simulating student learning for better AI tutoring · Microsoft Research

AI models can provably hide arbitrary capabilities · arXiv (Hacker News)

How an assessor could investigate alignment of unreleased AI models · METR

How an assessor could investigate control and monitoring of internal AI agents · METR

How an assessor could investigate evidence of acceleration from the automation of AI R&D · METR

Redesigning source control at Anthropic · Simon Willison

How the Claude Code team does code review · Hacker News: Claude Code

Quoting Addy Osmani, Morgan Westlee Lunt, and Qing Wang · Simon Willison

Postgres 19 delayed after AI reviewers find more bugs than humans can fix · Hacker News: agents

How Anthropic's Dynamic Workflows Coordinate Claude Agents · Latent Space

Quoting Geoffrey Huntley · Simon Willison

How to engineer an AI harness for game development · Hacker News: agents

How to Rewrite Apache Spark UDFs for GPU Acceleration with AI Agents · NVIDIA Technical Blog

Optimizing Agent Reliability and Tool Calling Efficiency for Chip Design Using NVIDIA NeMo Gym · NVIDIA Technical Blog

web-edit 0.1a0 · Simon Willison

Profiling web-edit with Python 3.15: flame graphs and source heatmaps · Simon Willison

web-edit 0.1a1 · Simon Willison

Tree Appreciator · Simon Willison

Claude Code 2.1.300 · Claude Code changelog

Claude Code 2.1.301 · Claude Code changelog

Claude Code 2.1.302 · Claude Code changelog

Cline v4.1.24 · Cline releases

Cline v4.1.25 · Cline releases

CLI v3.0.73 · Cline releases

Desktop v0.0.48 · Cline releases

SDK v0.0.95 · Cline releases

Codex CLI 0.163.0 · OpenAI Codex CLI releases

Release v0.64.0 · Gemini CLI releases

Release v0.65.0-preview.0 · Gemini CLI releases

Gemini CLI v0.65.0-nightly.20261013 · Gemini CLI releases

Release v0.66.0-nightly.20261017.ged96faedf · Gemini CLI releases

Origin App Marketplace · Cursor blog

Cloud Agent Canvases and Environment Improvements · Cursor blog

Linus Torvalds Working on a Hefty Linux Kernel Improvement with Claude Code · Hacker News: Claude Code

500B tokens later: Letting AI agents decompile a first-person shooter · Hacker News: agents

Five months treating bugs like patients and coding agents like a medical team · Hacker News: agents

docs: serve the TypeScript quickstart server with serveStdio (#3427) · MCP-määrittely commits

docs(community): rename Primitive Grouping IG to Progressive Disclosure WG (#3376) · MCP-määrittely commits

Show HN: SolidCraft, Open-Source Parametric CAD in Rust (With MCP Support) · Hacker News

Show HN: An experimental pattern-first DAW that supports MCP · Hacker News

Show HN: Edi Life OS: self-hosted life dashboard with an MCP server for AI · Hacker News: MCP

android/skills · GitHub trending -syöte

K-Dense-AI/scientific-agent-skills · GitHub trending -syöte (mshibanami/GitHubTrendingRSS)

add new category - Mods (#3184) · awesome-claude-code commits

Add resource: prismantis (#3186) · awesome-claude-code commits

Add resource: PLUR (#3152) · awesome-claude-code commits

Add resource: funes (#3179) · awesome-claude-code commits

Add resource: Mycelium (#3180) · awesome-claude-code commits

xerj-org/xerj · GitHub trending

stablyai/orca · GitHub trending -syöte (mshibanami/GitHubTrendingRSS)

A Terminal Protocol for Program Status (OSC 7501) · Hacker News: MCP

Talorgs: A self-hosted personal AI agent on Cloudflare's free tier · Hacker News: agents

Alishahryar1/free-claude-code · GitHub trending

ifixai-ai/iFixAI · GitHub trending -syöte (mshibanami/GitHubTrendingRSS)

How I manage persistent Claude Code loops: Spool · Hacker News: Claude Code

xtool-org/xtool · GitHub trending -syöte (mshibanami/GitHubTrendingRSS)

Show HN: Sygal: Convert PDFs, scans and web pages to Markdown, locally · Hacker News: MCP

Using AI to reverse engineer hardware · Hacker News: agents

AI-assisted PCB design for KiCad 10 · Hacker News: agents

Llama.cpp on a jailbroken PS5: run any GGUF model natively on the console · Hacker News: MCP

security-guidance: parse diff sections by header line (#6542) · Claude Code plugins/marketplace (virallinen)

security-guidance: count diff headers the parser could not read · Claude Code plugins/marketplace (virallinen)

security-guidance: never report a failed review as clean (#6375) · Claude Code plugins/marketplace (virallinen)

security-guidance: send ANTHROPIC_CUSTOM_HEADERS on the plugin's own API calls (#6378) · Claude Code plugins/marketplace (virallinen)

security-guidance: deliver commit and push findings on stderr (#6379) · Claude Code plugins/marketplace (virallinen)

security-guidance: report gateway routing and inner CLI stderr in review telemetry · Claude Code plugins/marketplace (virallinen)

Cline CLI v3.0.71 · Cline releases

Cline SDK v0.0.93 · Cline releases

Cline CLI v3.0.72 · Cline releases

Cline SDK v0.0.94 · Cline releases

n8n@2.44.0 · n8n releases

n8n@2.43.6 · n8n releases

n8n@2.44.2 · n8n releases

Toimitusjohtajan pankkitilin saldo lävähti koko firman nähtäville: Tekoäly jakoi tilitiedot työntekijöille · Tivi

TV4 postpones show after AI agent erases edited material · Hacker News: agents

Containment and monitoring failures are to blame for AI hacking reports: experts · Hacker News: agents

Ivanti Connect Secure -tuotteissa haavoittuvuutta käytetty hyväksi · Kyberturvallisuuskeskus

Cisco Secure Email Gateway -tuotteessa kriittinen hyväksikäytetty haavoittuvuus · Kyberturvallisuuskeskus

Syyskuun kybersää 2026 · Kyberturvallisuuskeskus