



Salkunhoitajien kvartaalikirje Q1 / 2026

Q1 / 2026: KOLMEN ILMOITUKSEN KVARTAALI

JOHDANTO

Q1 / 2026 oli vuosineljännes, jota leimasi kolme ilmoitusta – ilmoitusta, jotka ensi silmäyksellä vaikuttivat toisistaan riippumattomilta, mutta joiden yhteisvaikutus muutti tekoälyn käyttöönnoton logiikan peruuttamattomasti.

Me näimme sen.

Markkinat eivät nähneet. Sen sijaan ne jatkoivat tuttua kaavaansa: panikoivat tehokkuusloikista, vertasivat arvostuksia dotcom-kuplan huippuihin ja hakivat historiallisista ennätyksistä kattoja sille, mitä on mahdollista saavuttaa. Jokainen näistä reaktioista oli looginen vanhan maailman linssin läpi katsottuna. Jokainen oli mielestämme enemmän tai vähemmän väärässä.

Q4 / 2025-kirjeessämme puhuimme Jevonsin paradoksista – siitä, kuinka laskentakustannusten laskeminen ei hillitse kysyntää vaan räjäyttää sen. Tällä neljänneksellä saimme vahvistuksen toiselle, yhtä syvälliselle totuudelle: teknologia ei yksin ylitä kuilua markkinoille. Sen tekevät liittoutumat.

Geoffrey Mooren sanoin valtavirran asiakas ei osta teknologiaa. Hän ostaa ratkaisun. Tänä neljänneksellä tekoälyn suurimmat toimijat alkoivat rakentaa juuri sitä, eli kokonaista tuotetta, tavalla, jota markkinat eivät vielä osaa hinnoitella. Työmme on metsästä juuri näitä väärinymmärryksiä. Tällä neljänneksellä niitä riitti.

Tämä katsaus on Asilo Asset Management Oy:n tuottamaa markkinointimateriaalia. Tämä katsaus ei ole kehoitus merkitä, lunastaa tai vaihtaa rahasto-osuuksia. Sijoittajan tulee sijoituspäätöksiä tehdessään perustaa päätöksensä omaan arvioonsa sekä ottaa huomioon omat tavoitteensa ja taloudellinen tilanteensa. Sijoittajan ei tule perustaa sijoituspäätöstään tähän katsaukseen. Tätä katsausta laadittaessa on pyritty tietojen luotettavuuteen, Asilo Asset Management Oy ei voi taata tämän katsauksen sisältämien tietojen täydellisyyttä tai oikeellisuutta eikä vastaa sen sisältämien tietojen mahdollisista virheistä tai puutteista. Rahastosijoittamiseen liittyy aina taloudellinen riski. Riskit on esitelty tarkemmin rahaston avaintietoasiakirjassa ja rahastoesitteessä. Rahastoon tehdyn sijoituksen arvo voi nousta tai laskea ja sijoitetun pääoman voi menettää osittain tai kokonaan, eikä rahastolle asetettua tavoitettua välttämättä saavuteta. Historiallisen kehityksen perusteella ei voi tehdä luotettavaa arviota tulevista tuotoista. Sijoittajan tulee ennen sijoituspäätöksen tekemistä tutustua huolellisesti rahaston rahastoesitteeseen, avaintietoasiakirjaan ja sääntöihin, jotka ovat saatavissa GRIT Rahastoyhtiö Oy:stä ja tai Asilo Asset Management Oy:stä. Rahastoa hallinnoi GRIT Rahastoyhtiö Oy.

Elettiin vuotta 1999, ja Microsoftilla oli vastassaan ongelma. He olivat rakentaneet hallitsevan käyttöjärjestelmän henkilökohtaisille tietokoneille, ja nyt he olivat päättäneet valloittaa suuryritykset. Windows NT oli kykenevä. Exchange oli tehokas. Teknologia toimi. Mutta Microsoft tiesi, ettei pelkkä tuote itsessään riittänyt.

Ohjelmistojen myyminen monikansalliselle suuryritykselle tarkoitti, että jonkun oli myös suunniteltava arkkitehtuuri, hallittava siirtymä vanhoista järjestelmistä, koulutettava työntekijät ja otettava vastuu, kun jokin hajosi kello kahdelta aamuyöllä. Microsoft pystyi kirjoittamaan koodin. He eivät kuitenkaan voineet olla kymmenessä tuhannessa konesalissa samanaikaisesti. Asiakkaat, jotka halusivat Microsoftia, eivät halunneet vain Microsoftia – he halusivat kokonaisvaltaisen ratkaisun: ohjelmiston, käyttöönoton, jatkuvan tuen ja vastuunjaon. Toisin sanoen, asiakkaan silmissä Microsoft ei tarjonnut kokonaista tuotetta.

Voittaakseen Microsoft muodosti liittoutuman. Vuonna 2000 he perustivat yhdessä Accenturen kanssa Avanaden – yhteisyrityksen, jonka insinöörit jalkautuivat asiakasorganisaatioihin toteuttamaan Microsoftin teknologiaa ja toimittamaan lopputuloksia pelkkien lisenssien sijaan. Tämä liittoutuma teki tuotteesta kokonaisen. Vuosikymmenessä Avnade oli ottanut Microsoft-infrastruktuurin käyttöön maailman suurimmissa organisaatioissa, ja Microsoftin asema suuryrityksissä oli käytännössä horjumaton. Teknologia ei ollut muuttunut. Kokonaisuus oli.

KUILU JA SEN SELÄTTÄMINEN

Geoffrey Moore kuvaili tätä ongelmaa jo kolmekymmentä vuotta sitten teoksessaan *Crossing the Chasm*. Kuilu, joka koituu useimpien teknologiayritysten kohtaloksi, ei ole kuilu huonojen ja hyvien tuotteiden välillä. Se on kuilu sellaisen tuotteen välillä, joka toimii varhaisten omaksujien käsissä – heidän, jotka sietävät keskeneräisyyttä, integroivat puuttuvat palaset itse ja julistavat sanomaa kitkasta huolimatta – ja sellaisen tuotteen välillä, joka toimii valtavirran asiakkaalle, joka ei tällaista siedä.

Valtavirran asiakas ei osta teknologiaa. Hän ostaa ratkaisuja ongelmaansa. Jos teknologia ratkaisee vain osan ongelmasta, hän ei osta sitä – riippumatta siitä, kuinka hyvä tuo osa on. Tämä on se kuilu: etäisyys ydintuotteen tarjoaman ja asiakkaan todellisten tarpeiden välillä. Yritykset, jotka ylittävät tämän kuilun menestyksekkäästi, eivät lähes koskaan tee sitä pelkästään ydintuotetta parantamalla. Ne tekevät sen rakentamalla liittoutumia – kumppanuuksia sellaisten toimijoiden kanssa, jotka toimittavat puuttuvat osat, täydentävät kokonaistuotteen ja muuttavat teknologian ratkaisuksi, jonka valtavirran markkinat voivat ottaa käyttöön ilman sankaritekoja.

Tämä kaava on toistunut jokaisen suuren teknologiamurroksen kohdalla. Kun sijoittaja seuraa uuden teknologian nousua, kysymys ei kuulu ainoastaan, onko ydintuote hyvä. Kysymys kuuluu, rakentuuko sen ympärille parhailaan kokonaistuotetta – ja kuka sen rakentaa, ja kuinka nopeasti.

Q1 / 2026: LIITTOUMIEN AALTO SAAPUU

Tämän vuoden ensimmäisellä neljänneksellä vastaus saapui epätavallisen selkeänä ja nopeana. Maaliskuussa tuli yhden viikon sisään kolme ilmoitusta, jotka vaikuttivat toisistaan riippumattomilta.

1. Jeff Bezoksen kerrottiin keräävän 100 miljardia dollaria valmistavan teollisuuden yritysten hankkimiseksi ja tekoälyn käyttöönottamiseksi näissä yrityksissä.
2. OpenAI aloitti pitkälle edenneet neuvottelut Advent Internationalin, Bain Capitalin, Brookfieldin ja TPG:n kanssa yhteisyrityksen perustamiseksi.
3. Anthropic avasi samanaikaisesti rinnakkaiset neuvottelut Blackstonen, Hellman & Friedmanin ja Permiran kanssa. Tekoälylaboratoriot eivät olleet keräämässä pääomaa. Ne olivat rakentamassa kokonaisia tuotteita.

Luvut ovat niin massiivisia, että niiden äärelle on syytä pysähtyä. Nämä seitsemän yritystä, jotka on tähän mennessä mainittu näissä järjestelyissä, hallinnoivat yhteensä noin 2,7 biljoonan dollarin varallisuutta. Se on karkeasti kahdeksan kertaa koko Helsingin pörssin arvo. Kyse on vain osasta globaalia pääomasijoitusala, joka hallitsee yhteensä 13 biljoonaa dollaria. Tämä tarkoittaa, että tämän neljänneksen aikana nähdyt ilmoitukset ovat vasta alkusoitto, eivät koko sinfonia. Yksi liittoutuma neljän pääomasijoitusyhtiön kanssa vastaa satojen suuryrityssuhteiden avaamista samanaikaisesti – suhteiden, joissa pääomasijoitusyhtiön oma taloudellinen kannustin on nyt sidottu tekoälyn käyttöönoton onnistumiseen.

Bezosen liike täydentää kokonaiskuvaa. Project Prometheus – tekoäly-startup, jota Bezos oli mukana perustamassa loppuvuodesta 2025 – rakentaa malleja, jotka ymmärtävät ja simuloivat fyysistä maailmaa: materiaaleja, prototyyppejä rakennusta ja esituotannon suunnittelua. Fyysisessä tekoälyssä kuilu kokonaisvaltaisuuteen on huutava. Tekoälyä ei voi myydä ilmaillutai siruteollisuuteen ilman toimialaosaamista, luottamuksellisia suhteita insinööriimeihin ja vuosien kokemusta kenttätyöstä. Bezos ei osta tehtaita sijoitusharjoituksena. Hän ostaa kokonaistuotteen täydennyksen, jota yksikään ohjelmistotoimittaja ei pysty yksinään tarjoamaan. Portfolioyhtiöistä tulee samanaikaisesti toteutuskerros, näyteikkuna ja jakelukanava. Microsoft tarvitsi Accenturen. Prometheus tarvitsee tehdaslattian."

MIKSI TÄMÄ TULEE TOIMIMAAN?

Liittoutumamalli onnistuu silloin, kun kumppanit pystyvät näyttämään toisilleen jotain konkreettista.

Tekoälylaboratoriot eivät enää tarvitse monimutkaisia myyntipuheita pääomasijoittajien huomion saamiseksi; niille riittää yksi ainoa datapiste. Vain neljätoista kuukautta sitten Anthropicin annualisoitu liikevaihto oli miljardi dollaria. Tänään se on noussut räjähdysmäisesti 14 miljardiin. OpenAI:n liikevaihto ponnahti kahdesta miljardista yli 20 miljardiin dollariin vain kahdessa vuodessa, saavuttaen tämän rajan vuonna 2025. Brad Gerstner kutsui tätä ilmiötä All-In-podcastissa maaliskuussa "tekoälyn liikevaihdon atominhalkaisuhetkeksi". Tällaiselle kasvuvauhdille ei löydy esikuvaa liike-elämän historiasta. Ei ohjelmistoista. Ei kuluttajateknologiasta. Ei mistään. Tuote ei ole arvioinnissa. Se on tuotannossa maailmantalouden korkeimmilla tasoilla. Alkuunpaneva voima on saavutettu. Lumipallo vyöryy alas jyrkkää rinnettä: se kerää vauhtia omasta massastaan ja nielee kaiken tielleen osuvan.

Sitten he esittelevät Goldman Sachsin tapauksen:

Kuusi kuukautta kenttätyötä integroitujen insinöörien kanssa. Valittuna kaksi työprosessia juuri siksi, että ne olivat vastustaneet automatisointia vuosikymmeniä: kaupankäynnin ja transaktioiden kirjanpito sekä asiakkuuksien avaaminen. Molemmissa yhdistyvät dokumenttien analysointi, säädösten tulkinta ja poikkeustapausten hallinta tavalla, josta aiemmat ohjelmistosukupolvet eivät selviytyneet. Goldmanin tietohallintojohtaja Marco Argenti kuvaili testauksen paljastamaa tulosta: sama päättelykyky, joka kirjoittaa koodia, osaa myös navigoida KYC-tietokannat ja täsmäyttää transaktiot. Goldman yllättyi laajuudesta. Heidän ei olisi pitänyt yllättyä, mutta heidän yllätyksensä on kaupallisesti hyödyllistä: se tarkoittaa, että keskimääräisen pääomasijoitusyhtiön sisällä piilevä mahdollisuus on suurempi kuin kyseinen yhtiö tällä hetkellä itsekään uskoo.

Pääomasijoitusyhtiöille ei makseta ensisijaisesti liikevaihdon kasvattamisesta. Niille maksetaan käyttökattteen kasvattamisesta, kiertoaikojen lyhentämisestä ja sellaisen operatiivisen vivun luomisesta, joka oikeuttaa korkeat irtautumiskertoimet. Tämä liittoutuma tarjoaa pääomasijoitusyhtiöille äärimmäisen työkalun marginaalien laajentamiseen. Yhteisyritysrakenne ei ole ensisijaisesti pääoman keräämisen mekanismi. Se on toteutusväline. Se on tapa ottaa "Goldmanin pelikirja" ja monistaa se samanaikaisesti satoihin salkkuyhtiöihin, jolloin pääomasijoittajan oma taloudellinen kannustin on sidottu jokaisen käyttöönoton onnistumiseen. Tekoälylaboratorioille jokainen käyttöönotto luo asiakkaan, jonka vaihtokustannus kasvaa jokaisen integroidun työnkulun, jokaisen uudelleenrakennetun prosessin ja jokaisen syvemmälle liiketoiminnan todelliseen toimintaan upotetun mallin myötä.

Tämä yhteensopivuus on rakenteellinen, ei sattumanvarainen. Tekoälylaboratoriot tuovat teknologian ja näytöt. Pääomasijoitusyhtiöt tuovat salkkuyhtiöt ja mandaatin luoda arvoa. Bezos tuo pääsyn fyysiseen maailmaan, jota kumpikaan osapuoli ei saavuta yksin. Jokainen osapuoli toimittaa juuri sen, mitä muilta puuttuu – yhdessä ne muodostavat ehjän kokonaisuuden.

Tämä ei ole sattumaa. Kyseessä on kuilun ylittäminen, mutta ei tavalla, jota Geoffrey Moore kuvitteli. Moore vertasi haastetta Normandian maihinnousuun vuonna 1944: raakaan ja sankarilliseen rantautumiseen vihollistulen alla, missä selviytyminen itsessään oli epävarmaa. Tämä hetki tuntuu toisenlaiselta. Se ei muistuta niinkään D-Day-maihinnousua, vaan siirtymää vuoden 1943 Teheranista vuoden 1945 Jaltaan – jolloin "suuri kolmikko" istui saman pöydän ääreen ja alkoi jakaa sodanjälkeistä maailmaa. Se, mitä olemme todistamassa, ei ole epätoivoinen rynnäkkö, vaan uuden teollisen liittoutuman hiljainen muodostuminen – liittoutuman, joka muovaa maailmantalouden uuteen uskoon.

UUSI MALLI UUTEEN MARKKINATILANTEESEEN

Se, mitä olet juuri lukenut, on osoitus ajattelumallien verkoston voimasta. Se on historiallisten kaavojen tunnistamiseen pohjautuva analyysi. Se nojaa viitekehukseen, jota on myyty miljoonia kappaleita ja joka on koeteltu käytännössä vuosikymmenten teknologiamurroksissa. Se on ajattelumalli, joka tuottaa konkreettisia, hyödynnettäviä oivalluksia, vaikka ne eivät perustukaan numeroihin vaan ymmärrykseen.

Mieti, mitä koko tuotekehys tarkoittaa kysynnän kannalta. Jokaisesta tekoälyagentin käyttöön ottavasta yrityksestä tulee jatkuva ja toistuva päättelykapasiteetin (eli inferenssin), muistikaistan ja verkkokapasiteetin suurkuluttaja. Vaihtokustannukset kertyvät nopeasti: esimerkiksi Goldman Sachsin kauppojen täsmäytysprosessiin integroitua mallia ei vaihdeta uuteen neljännesvuosittaisen hankintasyklin mukana. Kun liittoutumat on kerran muodostettu, ne eivät

ainoastaan nopeuta teknologian omaksumista – ne tekevät siitä pitkäkestoisen. Kysynnän seuraus on rakenteellinen, ei syklinen, ja se kumuloituu.

Voisi olettaa, että markkinat lukevat tilanteen näin laajemminkin. Loppujen lopuksi *Crossing the Chasm* ei ole mikään hämärä akateeminen julkaisu. Jokainen liikkeenjohdon konsultti, pääomasijoittaja ja teknologia-alan strategioista vastaava johtaja on lukenut sen viimeisen kolmenkymmenen vuoden aikana. Mutta markkinat eivät lue tilannetta näin. Merkittävä osa institutionaalaisesta analyysistä nojaa yksinomaan takaperosiin kvantitatiivisiin mittareihin. Koviin lukuihin. Todennettavaan dataan. Skaalautuviin malleihin. Menetelmiin, jotka tuottavat täsmällisiltä kuulostavia ja siksi vakuuttavia argumentteja.

TARKASTELLAANPA, MILTÄ NUO ARGUMENTIT TODELLISUUDESSA NÄYTTÄVÄT.

Dotcom-kuplan huipulla vuonna 2000 Cisco Systemsin markkina-arvo saavutti 555 miljardia dollaria – se oli tuolloin korkein koskaan mitattu arvo millekään yritykselle koko maapallolla. Vuosien ajan tuosta luvusta tuli "järjettömän yltiöpäisyyden" (*irrational exuberance*) synonyymi. Tämän logiikan mukaan mikä tahansa tuota arvostustasoa lähestyvä yritys oli hakeutumassa vaaran vyöhykkeelle. Argumentti kuulostaa vakuuttavalta: verrataan nykyisiä arvostustasoja eliniän suurimman kuplan historialliseen huippuun ja päätellään, että kyseisen luvun läheisyys on merkki hypestä, kuplasta ja järjettömistä arvostuksista.

Ongelma on siinä, että nykyään on kymmeniä yrityksiä, joiden markkina-arvo saa Cison dotcom-huipun näyttämään kääpiömäiseltä. Pelkästään Nvidia on ylittänyt 4 biljoonan dollarin rajan. Luku, joka kerran määritti inhimillisen taloudellisen mielikuvituksen äärirajat, on nykyään S&P 500 -indeksin keskivaiheen lukema. Argumentti ei ollut väärä siksi, että laskuoppi olisi ollut virheellistä. Se oli väärä, koska siinä historiallista ääripäätä käsiteltiin pysyvänä kattona eikä pelkkänä väliaikaisena virstanpylväänä.

Sama virhe toistuu kasvuvauhdin analysoinnissa. Kuinka nopeasti yritys voi kasvaa? Kvantitatiivinen analytiikka tarkastelee historian nopeimpia esimerkkejä ja pitää niitä ehdottomina rajoina. Netflix saavutti 100 miljoonaa tilaajaa kymmenessä vuodessa – tuo katto määritti ylärajan sille, mitä kuluttajateknologian omaksumisessa voitiin saavuttaa. Sitten ChatGPT saavutti 100 miljoonaa käyttäjää 60 päivässä, nopeammin kuin mikään tuote historiassa. Sen jälkeen OpenClaw"sta tuli ohjelmistohistorian nopeimmin kasvanut avoimen lähdekoodin projekti, ylittäen muutamassa viikossa sen, mihin Linuxilta kului 30 vuotta. Mieti tuota eroa: 30 vuotta – ja muutama viikko! Joka kerta, kun malli kalibroitiin edellisen ennätyksen mukaan, seuraava kehitysaskel teki kalibroinnista vanhentuneen. Rajoitteena ei ollut teknologia, vaan analytiikan otoskoko.

Sama kaava toistuu avointen alustojen leviämisessä. Linux oli perinteinen esimerkki avoimen lähdekoodin yleistymisestä – vuosikymmeniä kestänyt matka harrastelijaprojektista yritysten infrastruktuurin standardiksi. Tuo matka loi ajatusmallin siitä, kuinka kauan avoimilta alustoilta vie "ylittää kuilu". OpenClaw ylitti sen viikoissa. Jensen Huang kutsui sitä maaliskuun 2026 GTC-tapahtumassa ihmishistorian nopeimmin kasvavaksi avoimen lähdekoodin projekti ja julisti, että jokainen yritys tarvitsee OpenClaw-strategian – sama kehys, jossa jokainen yritys tarvitsi aikoinaan Linux-strategian, on nyt tiivistetty vuosikymmenistä yhteen ainoaan tuloskauteen.

YKSITTÄISET KALIBROINTIVIRHEET KERTAUTUVAT INSTITUTIONAALISISSA MALLEISSA

Englannin keskuspankki ja IMF ovat vetäneet suoria vertauksia dot-com-kuplaan, viitaten arvostuskertoimiin ja velkatason kasvuun irrationaalisen yltäkylläisyyden todisteena. Goldman Sachsin analyytikot dokumentoivat hyperscalerien ottavan 121 miljardia dollaria uutta velkaa yhden vuoden aikana — 300 prosentin kasvu historiallisista normeista — ja kehystivät sen ylivenyttäytymisenä. Morgan Stanley arvioi tekoälyinfrastruktuuriin kuluvaan 3 biljoonaa dollaria vuoteen 2028 mennessä, mistä puolet on velkarahoitteista, ja esittää luvun varoituksena. Deutsche Bankin Jim Reid ennustaa OpenAI:lle 140 miljardin dollarin kumulatiivisia tappioita vuoteen 2029 mennessä ja päättelee, että toiminnan taloudellinen pohja on kestämaton.

Todellisuus on kuitenkin alirakentaminen, ei yllirakentaminen.

Tämän vuoden ensimmäisellä neljänneksellä viiden suurimman hyperscalerin investointiohjeistus kasvoi noin 150 miljardia dollaria yhden tuloskvartaalin aikana. Tämä ei ole luku, joka kumpuaa yliarvioinnista — se on luku, joka kumpuaa aliarvioinnista.

Vieläkin selvempi signaali tuli AWS:n toimitusjohtaja Matt Garmanilta helmikuussa: AWS ei ole poistanut käytöstä yhtään A100-palvelinta. Ei yhtäkään.

Ajattele tätä autoanalogian kautta. Jensen Huang lupasi, että hänen "nykyinen autonsa" eli Blackwell-grafiikkaprosessori kuluttaisi "5 litraa bensiiniä sadalla kilometrillä". Puolijohde- ja tekoälyalaan erikoistunut tutkimusyhtiö SemiAnalysis vei sen koeajolle ja mittasikin kulutukseksi vain 2,5 litraa satasella. Lupaa vähemmän, toimita enemmän!

Kuvittele nyt kokonainen autokanta. Uudet autot kuluttavat tuon 2,5 litraa, mutta ne vanhimmat vuoden 2020 "Datsunit" (eli Nvidia A100 -sirut) kuluttavat päivitetyn ohjelmistonkin kanssa yhä toistasataa litraa satasella — verrattuna päivitettyihin Blackwell -siruihin. Microsoftin talousjohtaja totesi tällä vuosineljänneksellä, että grafiikkaprosessoreita seisoo joutilaana varastoissa sähkönpuutteen vuoksi. Ja silti: yhtäkään tällaista "Datsunia" ei ole romutettu, vaikka elämme keskellä energiapullonkaulaa! Ja tässä vaiheessa emme ole vielä ottaneet kantaa loppuvuodesta 2026 tulevaan Nvidian uusimpaan, Vera Rubin -siruun jolloin "Datsunin" vertailukelpoinen kulutus olisi useamman sadan litran kokoluokkaa.

Tämä ei ole kupla. Näin tapahtuu, kun toimiala on rakentanut liian vähän kapasiteettia vuosien ajan ja yrittää nyt epätoivoisesti kuroa umpeen tuon kuilun. Operaattorit ajavat kuusi vuotta vanhaa laitteistoa täydellä teholla yksinkertaisesti siksi, ettei sen korvaamiseksi ole tarpeeksi infrastruktuuria — ei siksi, etteikö uutta teknologiaa olisi olemassa, vaan siksi, että sähköön, jäähdytykseen ja konesaltiloihin on investoitu liian vähän.

YKSITYISSIJOITTAJIEN KAIKUKAMMIO: TÄSMÄLLISYYTTÄ ILMAN YMMÄRRYSTÄ

Jos institutionaaliset mallit kärsivät kalibrointivirheistä, yksityissijoittajien markkinat kärsivät kontekstin puutteesta. Teknisesti ottaen tosia tietoja käytetään usein sellaisten johtopäätösten tekemiseen, jotka ovat loogisesti mahdottomia.

Otetaan esimerkiksi tuore reaktio Google Researchin "TurboQuant"-julkistukseen. Muutamassa tunnissa sosiaalinen media täyttyi tarinoista, joiden mukaan "tekoälykupla" oli puhjennut, koska uusi pakkausalgoritmi pienensi KV-välimuistin (KV-cache) muistivaatimuksia.

Arvostelukyvyyttömälle silmälle logiikka näyttää pätevältä: jos Google pystyy kutistamaan muistinkulutuksen kuudesosaan, tarvitsemme kuusi kertaa vähemmän siruja. Jos tarvitsemme vähemmän siruja, "muistiosakkeiden" on romahdettava. Se on suoraviivainen, selkeä ja täysin virheellinen päätelmä.



Ejaaz @cryptopunk7213 · Mar 25

wow google might've popped the ai bubble, memory stocks down massively today:

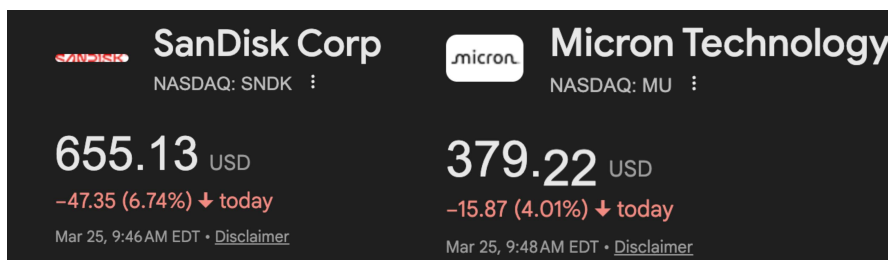
their new algorithm shrinks an AI model's memory by 6X WITHOUT reducing it's intelligence making it 8x faster with the SAME # of GPUs:

if this works - we don't need as many GPUs to train AI

- kv-cache is basically a model's short term memory. it gets massive pretty quickly = larger, slower, expensive ai
- google's algo compresses it to just 3-bits with ZERO loss in accuracy (usually models are like 32-bit)

the combined market cap of micron and sandisk is \$527 billion and im not even factoring in SK hynix and samsung

ai has driven up memory prices by 500%+ over the last few months - if google's algo scales then this might crash.



Company	Symbol	Price (USD)	Change (Today)
SanDisk Corp	NASDAQ: SNDK	655.13	-47.35 (6.74%) ↓
Micron Technology	NASDAQ: MU	379.22	-15.87 (4.01%) ↓

Mar 25, 9:46 AM EDT • Disclaimer

Mar 25, 9:48 AM EDT • Disclaimer



Google Research @GoogleResearch · Mar 24

Introducing TurboQuant: Our new compression algorithm that reduces LLM key-value cache memory by at least 6x and delivers up to 8x speedup, all with zero accuracy loss, redefining AI efficiency. Read the blog to learn how it achieves these results: ...

274 602 8.2K 1.6M

Tämä tietty "paniikki" perustuu kolmeen perustavanlaatuihin kalibrointivirheeseen:

1. **KV-välimuistiharha:** KV-välimuisti (Key-Value cache) ei ole tekoälyjärjestelmän "kokonaismuisti" — se on erityinen, dynaaminen muistin osajoukko, jota käytetään päättelyn aikana. Välimuistin jalanjäljen pienentäminen on uskomaton insinöörisuoritus,

mutta se ei poista massiivisia HBM-vaatimuksia (High Bandwidth Memory), joita tarvitaan itse mallin painoille. "Lyhytaikaisuusmuistin" pienentäminen ei yhtäkkiä tee "aivoista" pienempiä.

2. **Jevonsin paradoksi:** Tämä on teknologia-analyysin yleisin sokea piste. Kun resurssista tehdään tehokkaampi, sen kokonaiskulutus ei suinkaan laske – se kasvaa. Tekemällä tekoälystä kahdeksan kertaa nopeamman ja kuusi kertaa muistitehokkaamman Google ei ole vähentänyt GPU-prosessorien tarvetta – se on vain tehnyt mallien käyttöönotosta taloudellisesti kannattavaa kymmenen kertaa useammalle yritykselle. Tehokkuus laskee kynnystä, mikä puolestaan räjäyttää markkinoiden kokonaispotentiaalin.
3. **Luokitteluvirhe:** Postauksessa mainitaan SanDisk esimerkkinä "muistiosakkeesta", joka kärsii tilanteesta. SanDisk (jonka omistaa Western Digital) valmistaa pääasiassa NAND-flash-muistia kulutuselektroniikkaan ja SSD-levyihin. Tekoälybuumia taas ajavat HBM- ja DDR5-muistit, joita valmistavat SK Hynix, Micron ja Samsung. SanDiskin rankaiseminen kielimallien päättelylogiikan läpimurron vuoksi on kuin myisi Fordin osakkeet siksi, että joku keksii tehokkaamman suihkumoottorin.

Ja näitä voisimme luetella enemmänkin.

Me emme pelkää tällaisia otsikoita. Me toivotamme ne tervetulleiksi. Kaikki näkevät samat luvut, me näemme mitä ne tarkoittavat. Meillä on kyky erottaa Datsun Blackwellistä – ja välimuistioptimointi kysynnän romahduksesta.

Kun markkina reagoi "TurboQuantiin" myymällä koko puolijohdeteollisuuden toimitusketjun, syntyy massiivinen kuilu hinnan ja todellisuuden välille. Vähittäissijoittaja näkee "kuplan puhkeamisen" kun me taas näemme tehokkuusloikan, joka johtaa seuraavaan sataan miljoonaan käyttäjään.

Me metsästämme väärinymmärryksiä. Ja väärinymmärryksiä tässä ympäristössä riittää. Olemme vasta pääsemässä vauhtiin.

Kiitos luottamuksestanne,

Henri Blomster & Ernst Grönblom
Asilo Argon Salkunhoitajat