Toimivatko suuret robottien ”perusmallit” oikeasti paremmin? Huolellinen vastaus: hieman kyllä, eikä useimmista tutkimuksista voi päätellä sitä

Toyota Research Institute koulutti ”large behavior models” -malleja — robottipolitiikkoja, jotka esikoulutettiin noin 1,700 tunnilla monipuolista manipulaatiodataa — ja vertasi niitä alusta koulutettuihin yhden tehtävän politiikkoihin poikkeuksellisen huolellisesti: sokkoutetuilla, satunnaistetuilla, suurten otosten kokeilla (noin 1,800 todellista ja yli 47,000 simuloitua suoritusta) sekä oikeilla tilastollisilla testeillä. Tehtäväkohtaisen hienosäädön jälkeen suuret mallit pärjäsivät keskimäärin paremmin, tarvitsivat noin 3–5× vähemmän tehtäväkohtaista dataa ja olivat kestävämpiä olosuhteiden muuttuessa; suorituskyky parani tasaisesti esikoulutusdatan määrän kasvaessa. Ilman hienosäätöä ne eivät kuitenkaan johdonmukaisesti päihittäneet yhden tehtävän malleja, useat vaikutukset olivat niin pieniä, että ne näkyivät vasta suurilla otoksilla, ja arkinen datan normalisointivalinta vaikutti enemmän kuin arkkitehtuuri. Tulos on mitattua tukea robottien perusmallien suuntaukselle — ei yleiskäyttöinen robotti, ei zero-shot-generalisti eikä ”emergentti loikka” — sekä terävä varoitus siitä, että suuri osa robotiikasta saattaa mitata kohinaa.

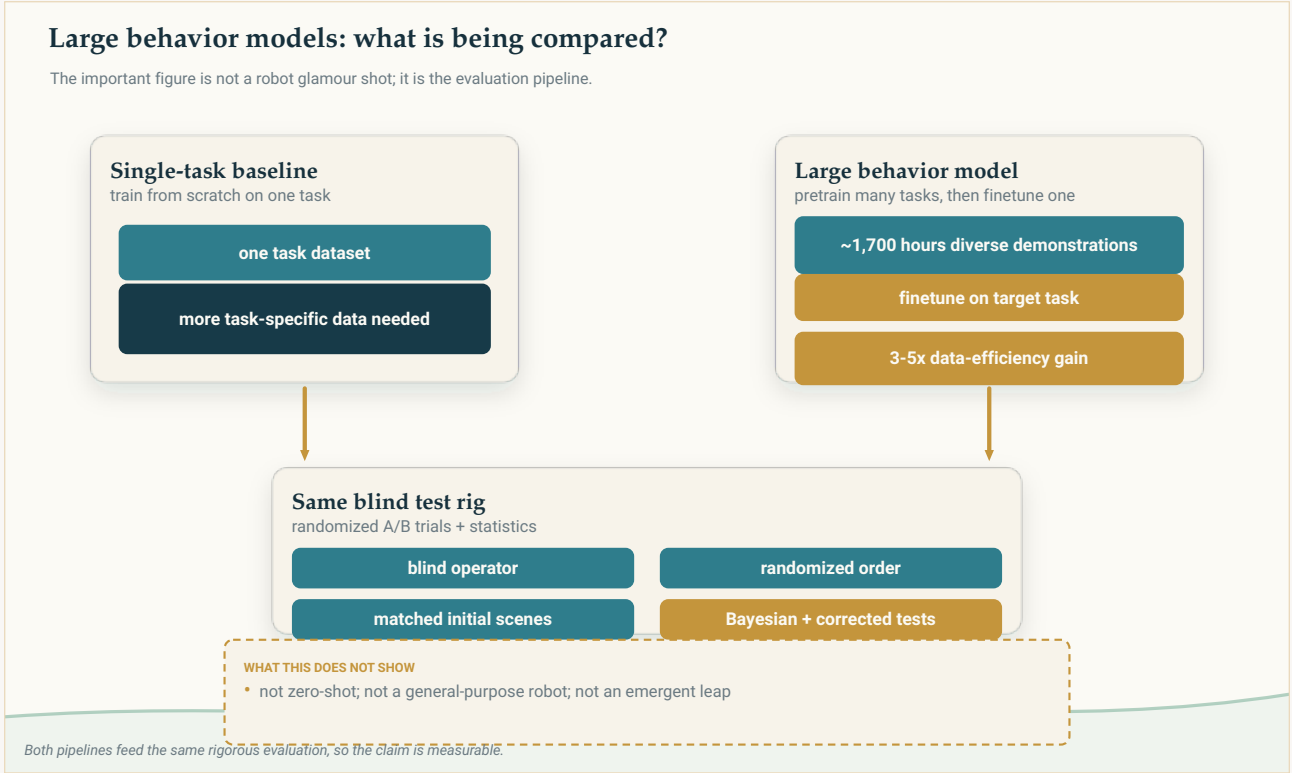
Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

Paper: *A Careful Examination of Large Behavior Models for Multitask Dexterous Manipulation*, Toyota Research Institute Large Behavior Model Team — J. Barreiros, A. Beaulieu, et al.; senior authors incl. R. Ambrus, B. Burchfiel, S. Feng, H. Kress-Gazit (Cornell), R. Tedrake, Science Robotics (2026); preprint arXiv:2507.05331 · DOI: 10.1126/scirobotics.aea6201

Robotiikkapaperi, jonka todellinen aihe on rehellisyys

Robotiikka elää ”perusmallien” huumaa. Kielestä ja kuvatekoälystä lainattu ajatus on houkutteleva: sen sijaan että robotti koulutettaisiin yksi tehtävä kerrallaan, koulutetaan yksi suuri **”large behavior model” (LBM)** valtavalla ja monipuolisella demonstraatioaineistolla ja saadaan järjestelmä, joka osaa laajasti erilaisia asioita ja mukautuu nopeasti. Innostus — ja investoinnit — ovat valtavia. Otsikko kirjoittaa melkein itse itsensä: *yleiskäyttöiset robottiaivot ovat täällä.*

Toyota Research Instituten paperi on kiinnostava juuri siksi, ettei se kirjoita tätä otsikkoa. Sen todellinen panos ei ole näyttävämpi robotti, vaan tiukka tarkastelu petollisen yksinkertaiseen kysymykseen — *toimiiko suuren mallin lähestymistapa oikeasti paremmin, ja mistä edes tietäisimme sen?* — tilastollisella huolellisuudella, joka on kirjoittajien mukaan alalla epätavallista. Tulos on aidosti hyödyllinen ”kyllä, mutta”, ja samalla varoitus siitä, että suuri osa robotiikasta saattaa mitata kohinaa.



Molemmat lähestymistavat testataan samalla sokkoutetulla ja satunnaistetulla arvioinnilla. Paperin väite ei ole, että robottien perusmallit olisivat zero-shot-generalisteja, vaan että esikoulutus voi parantaa datatehokkuutta ja robustiutta, kun vaikutus mitataan huolellisesti.

Original The Clean Paper diagram CC BY 4.0

Mitä kirjoittajat tekivät

He rakensivat LBM-malleja täsmällisessä merkityksessä: **diffuusiopohjaisia visuomotorisia politiikkoja** (Diffusion Transformer, joka lukee kamerakuvia, lyhyen kieliohjeen ja robotin omien nivelten

asennot ja tuottaa lyhyitä moottorikomentojen sarjoja taajuudella 10 Hz). Mallit **esikoulutettiin noin 1,700 tunnilla** robottidemonstraatioita — yli 500 erilaista talon sisällä kerättyä tehtävää sekä julkisia aineistoja — ja sitten **hienosäädettiin** yksittäisiin tehtäviin. Vertailukohtana koko paperissa on **yhden tehtävän politiikka, joka koulutettiin alusta** vain kyseisen tehtävän datalla.

Paperin ydin on kuitenkin *arviointi*, jota kirjoittajat käsittelevät pääasiallisena tuloksena. Jotta he eivät huijaisi itseään, he käyttivät:

- **Sokkoutettua, satunnaistettua A/B-testausta** oikeassa maailmassa — robottia käyttävä ihminen ei tiennyt, mitä politiikkaa kulloinkin testattiin, ja järjestys satunnaistettiin.
- **Hallittuja, toistettavia alkuolosuhteita** — käyttäjät sovittivat tilanteen kuva-overlayhin ennen jokaista koetta.
- **Suuria koemääriä** — 50 todellisen maailman suoritusta tehtävää, politiikkaa ja olosuhdetta kohti; simulaatiossa 200 tehtävää kohti. Yhteensä noin **1,800 sokkoutettua todellisen maailman suoritusta ja yli 47,000 simulaatiosuoritusta**.
- **Kunnollisia tilastoja** — Bayesilaisia onnistumistodennäköisyyden arvioita ja pareittaisia hypoteesitestejä monivertailukorjauksineen sen sijaan, että vain katsottaisiin päällekkäisiä virhepalkkeja. He tekivät jopa laadunvarmistuskierroksen neljännekselle ihmisten pisteyttämistä suorituksista mittaakseen pisteytysvirhettä.

[video pending]

Rinnakkainen vertailu malleista kattamassa aamiaispöytää: vasemmalla yhden tehtävän perustaso ja oikealla LBM. Molemmat videot toistuvat nopeudella 1x. Tämä on yksi arvioitu tehtävä, ei todiste yleiskäyttöisestä autonomiasta.

Tämä koneisto on paperin varsinainen pointti. Koko työ on argumentti siitä, ettei ilman sitä voi erottaa todellista parannusta sattumasta.

Mitä he löysivät

Hienosäädetyt suuret mallit päihittivät alusta koulutetut yhden tehtävän mallit — keskimäärin. Kun tehtävät yhdistettiin, esikoulutettu ja tehtäväkohtaisesti hienosäädetty LBM suoriutui luotettavasti paremmin kuin samalle tehtävälle alusta koulutettu politiikka sekä simulaatiossa että oikeassa maailmassa, ja ero oli tilastollisesti merkitsevä. Yksittäisissä tehtävissä hienosäädetty LBM oli tilastollisesti yhtä hyvä tai parempi lähes aina (3/3 todellisen maailman tehtävää, 15/16 simulaatiotehtävää).

Suurin ja selkein voitto on datatehokkuus. Hienosäädetty LBM saavutti alusta koulutetun mallin tasoisen suorituskyvyn noin **3–5× vähemmällä tehtäväkohtaisella datalla**. Yhdessä todellisen maailman tehtävässä (aamiaispöydän kattaminen) LBM, joka hienosäädettiin vain **15%**:lla demonstraatioista, päihitti alusta koulutetun politiikan, joka käytti **100%** niistä.

Esikoulutus auttaa eniten olosuhteiden muuttuessa. Kun testiympäristöä muutettiin tarkoituksella koulutusolosuhteista poikkeavaksi (”distribution shift”), hienosäädetyn LBM:n etu *kasvoi*. Yhdessä

simulaatiosarjassa se päihitti alusta koulutetun mallin tilastollisesti 3:ssa 16 tehtävästä normaaliolosuhteissa mutta 10:ssä 16 tehtävästä jakaumasiirtymän aikana. Koska todellinen käyttöympäristö poikkeaa aina jonkin verran koulutuksesta, tämä kestävyys on ehkä käytännössä tärkein havainto.

Lisää esikoulutusdataa auttoi tasaisesti. Suorituskyky parani vakaasti esikoulutusdatan määrän kasvaessa — testatuilla skaaloilla **ei nähty äkillistä hyppyä tai ”emergenttiä” loikkaa**. Hyödyllistä, ennustettavaa ja epädramaattista.

Mutta ilman hienosäätöä toimivan generalistin tarina ei pitänyt. Esikoulutettu LBM, jota käytettiin *zero-shot*-tilassa eli ilman tehtäväkohtaista hienosäätöä, *ei* johdonmukaisesti päihittänyt yhden tehtävän politiikkoja. Yksi verkko pystyi tekemään monta tehtävää, mutta unelma ”anna vain sanallinen ohje” ei toteutunut tässä; kirjoittajat liittävät osan ongelmasta pienen kielienkooderinsa haurauteen.

Ja hyödyt olivat niin pieniä, että ne olisi helppo ohittaa — tai kuvitella. Monet vaikutuksista tulivat näkyviin vasta tavallista suuremmilla otoksilla ja huolellisilla testeillä. Kirjoittajat toteavat suoraan, että vaikutusten koon ja kohinan vuoksi **on merkittävä riski, että monet robotiikkapaperit mittaavat tilastollista kohinaa.** He havaitsivat myös, että arkinen valinta — miten data *normalisoidaan* — vaikutti tuloksiin enemmän kuin arkkitehtuurimuutokset, ja esikoulutuksen normalisointi-**bugi** löytyi vasta *arviointien päätyttyä*.

Mitä tämä todennäköisesti tarkoittaa

Puolustettava tulkinta on, että laajamittainen esikoulutus monipuolisella robottidatalla on todellinen ja hyödyllinen osatekijä: se vähentää uuden tehtävän tarvitseman datan määrää ja tekee politiikoista kestävämpiä, kun maailma ei vastaa koulutusolosuhteita. Se tukee aidosti suuntaa, johon ala panostaa. Hyödyt ovat kuitenkin *maltillisia ja ehdollisia* — ne näkyvät pääasiassa hienosäädön jälkeen ja selkeimmin koontituloksissa sekä stressiolosuhteissa — eivätkä merkitse valmiin yleisrobotin saapumista.

Hiljaisempi mutta tärkeämpi merkitys on metodologinen. Paperi toimii käytännössä mittatikkuna: se näyttää, kuinka paljon näyttöä tarvitaan, jotta robottipolitiikasta tehtyyn väitteeseen voi luottaa, ja vihjaa, että suuri osa julkaistusta innostuksesta lepää liian vähäisen näytön varassa. Ala tarvitsee tällaista korjausliikettä enemmän kuin vielä yhden mallin.

Mitä tämä *ei* osoita

- Se **ei ole yleiskäyttöinen robotti.** Hyödyt osoitetaan tietylle arkkitehtuurille (diffuusiopolitiikoille), jotka hienosäädetään tehtäväkohtaisesti hallituissa oloissa teleoperaatiodemonstratioista — ei autonomiselle robotille, joka tekee käskystä mielivaltaisia uusia tehtäviä.
- Se **ei** vahvista **zero-shot**-käyttöä. Ilman hienosäätöä suuri malli ei johdonmukaisesti päihittänyt yhden tehtävän perustasoja.

- Se **ei** ole näyttöä **”emergentistä loikasta”.** Skaalaus paransi tuloksia tasaisesti; tässä ei ole epäjatkuvuutta, joka tukisi tarinaa ”ja sitten se yhtäkkiä osasi”.
- Luvut ovat **suhteellisia ja laboratorioon sidottuja.** Absoluuttiset onnistumisprosentit säädettiin tarkoituksella lähelle ~50%:a, jotta vertailut olisivat herkkiä; ne eivät mittaa todellisen maailman luotettavuutta, ja kyse on yhdestä arkkitehtuurista yhdessä laboratoriossa.
- Se **ei** ratkaise, **miksi** jokin politiikka onnistuu tai epäonnistuu, ja paperi raportoi useita yksittäisiä tehtäviä, joissa suuri malli pärjäsi *huonommin,* selittämättä syytä.
- Se ei kerro **turvallisuudesta, autonomiasta tai käyttöönotosta** arviointijärjestelyn ulkopuolella.

Kuinka vahvaa näyttö on?

Keskeisille vertailuväitteille — hienosäädetyt LBM:t päihittävät alusta koulutetut perustasot koontituloksissa, tarvitsevat moninkertaisesti vähemmän dataa ja ovat kestävämpiä jakaumasiirtymässä — näyttö on vahvaa *ja* poikkeuksellisen hyvin kontrolloitua: sokkoutettua, satunnaistettua, suurilla otoksilla tehtyä ja tilastollisesti testattua, ja lisäksi pisteytykselle tehtiin laadunvarmistus. Tässä harvinaisessa tapauksessa metodologia on riittävän vahva, jotta tärkeimmät johtopäätökset voi ottaa lähes sellaisinaan.

Rehelliset varaukset ovat kirjoittajien itsensä nostamia. Virhepalkit kuvaavat *arvioinnin* satunnaisuutta, mutta **eivät koulutuksen satunnaisuutta** — saman mallin kahdesta koulutuskerrasta voi tulla merkittävästi erilaiset politiikat, eikä tätä vaihtelua ole tilastoissa. Todellisen maailman tehtävissä oli 50 koetta kutakin, mikä riittää keskisuuriin vaikutuksiin mutta voi ohittaa pienet. Kielikonditionointi käytti vaatimatonta enkooderia, joten väitteet ”kerro vain robotille mitä tehdä” voivat näyttää erilaisilta suuremmissa järjestelmissä. Lisäksi paperi kertoo avoimesti jälkikäteen löytyneestä normalisointibugista. Mikään näistä ei kaada päätuloksia, mutta ne ovat juuri niitä asioita, jotka paperin mukaan ala usein lakaisee sivuun.

Lähteistä vielä yksi huomio samassa hengessä: tämä selitys perustuu kirjoittajien preprintiin. Emme saaneet journalissa julkaistua versiota haettua, joten emme ole tarkistaneet, muuttuiko teksti preprintin ja julkaistun version välillä.

Miksi tämä on tärkeää

”Robottien perusmallit” on ilmaus, joka melkein kutsuu ylilyönteihin, ja tutkimuksen voi lukea väärin kumpaankin suuntaan — voitonriemuisena *”se toimii!”* tai vähättelevänä *”se on ylihypetetty.”* Täsmällinen tulkinta on hyödyllisempi kuin kumpikaan: monipuolisella datalla tehty esikoulutus tuo todellisia, mitattavia mutta maltillisia hyötyjä — ennen kaikkea vähemmän dataa tehtävää kohti ja enemmän kestävyyttä — ja kehitys paranee ennustettavasti skaalan mukana.

Syvempi merkitys on siinä, että paperi kohdistaa tarkkuutensa omaan alaansa. Osoittamalla, että todelliset vaikutukset ovat niin pieniä, että ne katoavat huolimattomassa arvioinnissa, ja että tylsä valinta kuten datan normalisointi voi merkitä enemmän kuin kekseliäs uusi arkkitehtuuri, se perustelee,

että suuri osa robottien oppimisen edistyksestä tarvitsee tukevampaa mittaamista ennen kuin siihen voi uskoa. Paperi, joka käyttää uskottavuutensa tuloksen ja toiveen välisen rajan vartioimiseen, tekee jotain harvinaisempaa ja arvokkaampaa kuin tulostaulukon voittaminen.

Tiivistelmä

Toyota Research Instituten tutkijat kouluttivat ”large behavior models” -malleja — diffuusiopohjaisia robottipolitiikkoja, jotka esikoulutettiin ~1,700 tunnilla monipuolista manipulaatiodataa — ja vertasivat niitä alusta koulutettuihin yhden tehtävän politiikkoihin poikkeuksellisen tiukalla protokollalla: sokkoutetulla, satunnaistetulla, suurten otosten arvioinnilla (≈1,800 todellista ja 47,000+ simulaatiokoetta) sekä oikeilla tilastollisilla testeillä. Tehtäväkohtaisen hienosäädön jälkeen suuret mallit pärjäsivät koontituloksissa luotettavasti paremmin, tarvitsivat noin **3–5× vähemmän tehtäväkohtaista dataa** ja olivat **kestävämpiä olosuhteiden muuttuessa**, ja suorituskyky parani tasaisesti esikoulutusdatan kasvaessa. **Ilman hienosäätöä** ne eivät kuitenkaan johdonmukaisesti päihittäneet yhden tehtävän malleja, useat vaikutukset olivat niin pieniä, että vasta suuret otokset paljastivat ne, ja arkinen datan normalisointivalinta vaikutti enemmän kuin arkkitehtuuri. Tulos on vahvaa ja mitattua tukea robottien perusmallien suunnalle — **ei** yleiskäyttöinen robotti, ei zero-shot-generalisti eikä ”emergentti loikka” — sekä terävä varoitus siitä, että suuri osa robotiikasta saattaa mitata kohinaa.

Realistinen arvio

Mitä paperi osoittaa: Tiukalla, sokkoutetulla ja tilastollisesti riittävän tehokkaalla arvioinnilla (≈1,800 todellisen maailman + 47,000+ simulaatiosuoritusta) monitehtäväisesti esikoulutetut ja sitten hienosäädetyt diffuusiopolitiikat (LBM:t) päihittävät koontituloksissa alusta koulutetut yhden tehtävän politiikat, saavuttavat vastaavan suorituskyvyn ~3–5× vähemmällä tehtäväkohtaisella datalla ja kestävät jakaumasiirtymää paremmin; suorituskyky kasvaa tasaisesti esikoulutusdatan mukana.

Mikä on uskottavaa mutta ei osoitettu: Että hyödyt siirtyvät paljon suurempiin vision-language-action-malleihin (heidän kielienkooderinsa oli pieni); että tasainen skaalaus jatkuu testatun datamäärän ulkopuolelle.

Mitä se ei osoita: Yleiskäyttöistä tai zero-shot-robottia (ei hienosäätöä → ei johdonmukaista etua); mitään ”emergenttiä” kyvykkyyshyppyä; selityksiä tiettyjen tehtävien epäonnistumisille; todellisen maailman absoluuttista luotettavuutta (onnistumisprosentit säädettiin herkkyyden vuoksi lähelle 50%:a); mitään turvallisuudesta tai autonomisesta käyttöönotosta.

Tärkeimmät rajoitukset: Tilastot kattavat arvioinnin satunnaisuuden mutta eivät koulutuskertojen satunnaisuutta; 50 todellisen maailman koetta tehtävää kohti voi ohittaa pienet vaikutukset; yksi arkkitehtuuri ja yksi laboratorio; vaatimaton kielienkooderi; datan normalisointibugi löytyi arvioinnin jälkeen; analyysi perustuu preprintiin (julkaistua versiota ei tarkistettu).

Kuinka paljon luottamusta tavallisella lukijalla pitäisi olla? Paljon siihen, että monitehtäväinen esikoulutus ja hienosäätö tuottavat todellisia, maltillisia hyötyjä — erityisesti datatehokkuudessa ja robustiudessa — ja että ne mitattiin poikkeuksellisen huolellisesti. Paljon siihen, että tämä **ei** ole yleiskäyttöinen tai zero-shot-robotti eikä **emergentti loikka**. Kohtalaisesti siihen, kuinka pitkälle hyödyt skaalautuvat suurempiin malleihin. Ja kirjoittajien oma varoitus kannattaa ottaa vakavasti: alan vaikutukset ovat niin pieniä, että alitehoiset tutkimukset saattavat raportoida kohinaa. Sopiva asenne on mitattu optimismi lähestymistapaa kohtaan ja terve skeptisyys robotti-AI-tuloksia kohtaan, joilla ei ole tällaista tilastollista tukea.

Lähteet

arXiv:2507.05331
https://arxiv.org/abs/2507.05331

Peer-reviewed version (not retrieved) — Science Robotics, 10.1126/scirobotics.aea6201
https://doi.org/10.1126/scirobotics.aea6201

Project page
https://toyotaresearchinstitute.github.io/lbm1/