The Clean Paper

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

Vélmennagrein sem fjallar í raun um heiðarleika

Vélmennarannsóknin prófar stór hegðunarlíkön sem forþjálfuð eru á fjölbreyttum gögnum og spyr hversu mikið þau flytja milli verkefna, vélmenna og aðstæðna. Niðurstöðurnar sýna gagnlegan flutning en líka skýrar bilanir við dreifingarbreytingu og engin sönnun á almennu zero-shot vélmenni. Hún er fyrst og fremst varkár mat á því hvar foundation-model hugmyndin virkar og hvar ekki.

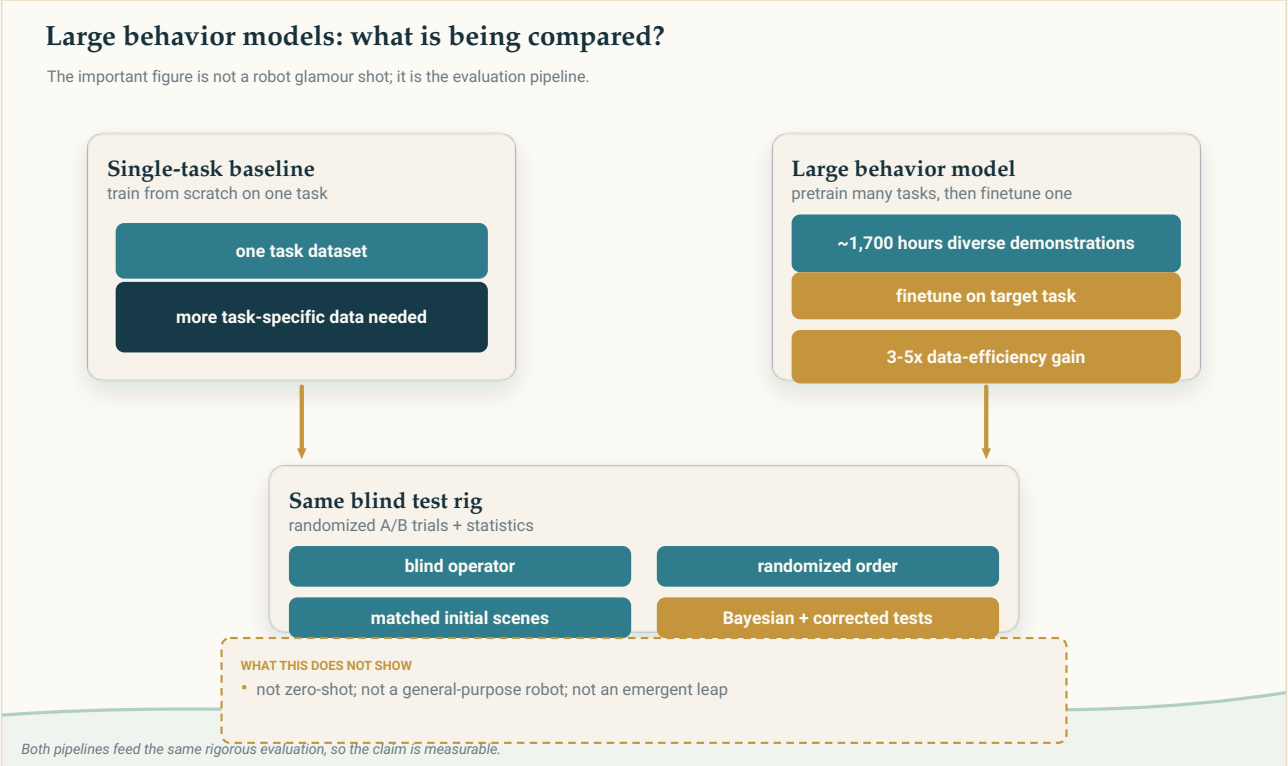
Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

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

Vélmennagrein sem fjallar í raun um heiðarleika

Vélmennafræði er mitt í bylgju „grunnlíkana“. Hugmyndin, fengin að láni frá tungumála- og myndgervigreind, er heillandi: í stað þess að þjálfa vélmenni eitt verkefni í einu, þjálfa eitt stórt **„large behavior model“ (LBM)** á gríðarstóru og fjölbreyttu safni sýnidæma og fá kerfi sem getur margt og aðlagast fljótt. Spennan — og fjárfestingin — er gífurleg. Fyrirsögnin skrifar sig sjálf: *alhliða vélmennagreind er komin.*

Grein Toyota Research Institute er áhugaverð einmitt vegna þess að hún neitar að skrifa þá fyrirsögn. Meginframlagið er ekki glæsilegra vélmenni. Það er harðsnúin athugun á blekkjandi einfaldri spurningu — *virkar stór-líkanaaðferðin raunverulega betur, og hvernig gætum við vitað það?* — með tölfræðilegri nákvæmni sem höfundarnir segja sjálfir óvenjulega á þessu sviði. Niðurstaðan er gagnlegt „já, en“ og viðvörun um að stór hluti vélmennafræðinnar kunni að vera að mæla hávaða.



Báðar aðferðir fara í gegnum sömu blindu, slembiröðuðu prófunina. Greinin heldur ekki fram að vélmennagrunnlíkön séu zero-shot alhæfingarkerfi; hún sýnir að forþjálfun getur aukið gagnanýtni og þol gegn aðstæðubreytingum þegar það er mælt vandlega.

Hvað gerðu höfundarnir?

Þeir smíðuðu LBM í mjög ákveðinni merkingu: **diffusion-byggð sjón-hreyfistefna** — Diffusion Transformer sem les myndavélar, stutta tungumálafyrirmæli og liðstöðu vélmennisins og skilar

stuttum runum hreyfiskipana við 10 Hz. Kerfin voru **forþjálfuð á um 1.700 klukkustundum** af sýnidæmum frá vélmennum — yfir 500 ólík verkefni sem safnað var innanhúss ásamt opinberum gagnasöfnum — og síðan **fínþjálfuð** á einstökum verkefnum. Samanburðurinn er allan tímann við **einstaklingsverkefnastefnu sem var þjálfuð frá grunni** aðeins á gögnum þess verkefnis.

Kjarni greinarinnar er þó *matferlið*, sem höfundarnir meðhöndla næstum sem aðalniðurstöðu. Til að blekkja sig ekki notuðu þeir:

- **Blinda, slembiröðaða A/B-prófun** í raunheimi — sá sem keyrði vélmennið vissi ekki hvaða stefna var prófuð og röðin var slembuð.
- **Stýrðar, endurtekningarhæfar upphafsaðstæður** — rekstraraðili stillti vettvanginn eftir myndyfirlagi fyrir hverja prófun.
- **Stóran fjölda keyrslna** — 50 raunheimskeyrslur á hvert verkefni, stefnu og ástand; 200 á verkefni í hermun. Samtals um **1.800 blindar raunheimskeyrslur og yfir 47.000 hermunarkeyrslur**.
- **Rétta tölfræði** — Bayesískt mat á líkum á árangri og pörunarpróf með leiðréttingum fyrir mörg samanburðarpróf, fremur en að horfa bara á skarandi skekkjustikur. Þeir gerðu meira að segja gæðapróf á fjórðungi mannaeinkunnanna til að meta skekkju í stigagjöf.

[video pending]

Samanburður hlið við hlið á líkönum sem leggja á morgunverðarborð: til vinstri grunnlíkan fyrir eitt verkefni, til hægri LBM. Bæði myndskeið eru á 1× hraða. Þetta er eitt metið verkefni, ekki sönnun á alhliða sjálfstæði.

Þetta matferli er sjálft punkturinn. Öll greinin er rök fyrir því að án þess sé ekki hægt að greina raunverulega framför frá heppni.

Hvað fundu þeir?

Fínþjálfuð stór líkön voru betri að meðaltali en eins-verkefnislíkön frá grunni. Þegar verkefni voru lögð saman stóð LBM sem hafði verið forþjálfað og síðan fínþjálfað á verkefninu sig áreiðanlega betur en stefna sem hafði aðeins verið þjálfuð frá grunni á sama verkefni, bæði í hermun og raunheimi, og munurinn var tölfræðilega marktækur. Á einstökum verkefnum var fínþjálfaða LBM-ið tölfræðilega jafn gott eða betra í nær öllum tilvikum (3/3 raunheimsverkefni og 15/16 hermunarverkefni).

Stærsti og skýrasti ávinningurinn var gagnanýtni. Fínþjálfað LBM náði frammistöðu sambærilegri við frá-grunni líkan með um **3–5× minna verkefnissértæku gögnum**. Í einu raunheimsverkefni, að leggja á morgunverðarborð, vann LBM sem var fínþjálfað á aðeins **15%** sýnidæmanna stefnu sem hafði verið þjálfuð frá grunni á **100%** þeirra.

Forþjálfun hjálpaði mest þegar aðstæður breyttust. Þegar prófunarumhverfinu var vísvitandi breytt frá þjálfunaraðstæðum — *distribution shift* — varð forskot fínþjálfaða LBM stærra. Í einni hermunarröð vann það frá-grunni líkön tölfræðilega á 3 af 16 verkefnum við venjulegar aðstæður en 10 af 16 þegar dreifingin færðist. Þar sem raunveruleg notkun víkur alltaf að einhverju leyti frá þjálfunaraðstæðum er þessi seigla hugsanlega mikilvægasta hagnýta niðurstaðan.

Meiri forþjálfunargögn hjálpuðu á sléttan hátt. Frammistaða jókst jafnt og þétt þegar forþjálfunargögnum var bætt við — **án skyndilegs stökks eða „emergent“ hæfileikabyltingar** á því stærðarbili sem var prófað. Gagnlegt, fyrirsjáanlegt og ódramatískt.

En sagan um alhæfingu án fínþjálfunar stóðst ekki. Forþjálfað LBM notað *zero-shot* — án verkefnissértækrar fínþjálfunar — vann ekki eins-verkefnislíkön stöðugt. Eitt net gat framkvæmið mörg verkefni, en draumurinn um „segðu því bara hvað það á að gera“ fékk ekki staðfestingu hér. Höfundarnir rekja hluta þess til smás og viðkvæms tungumálakóðara kerfisins.

Og ávinningurinn var nógu lítill til að auðvelt væri að missa af honum — eða sjá hann þar sem hann er ekki. Mörg áhrif urðu aðeins greinileg með stærri úrtökum og vandaðri tölfræði en venjulega. Höfundarnir segja hreint út að miðað við áhrifastærðirnar og hávaðann sé **veruleg hætta á að margar vélmennagreinar séu að mæla tölfræðilegan hávaða**. Þeir fundu einnig að hversdagslegt atriði — hvernig gögnin voru *normaliseruð* — hafði meiri áhrif en breytingar á byggingu líkansins, og normaliserunar**villa** í forþjálfun fannst aðeins *eftir* að matinu lauk.

Hvað þýðir þetta líklega?

Varnanlega túlkunin er að stórfelld forþjálfun á fjölbreyttum vélmennagögnum sé raunverulega gagnleg: hún dregur úr því gagnamagni sem þarf fyrir nýtt verkefni og gerir stefnu seigari þegar heimurinn passar ekki við þjálfunina. Það styður stefnuna sem fræðigreinin veðjar á. En ávinningurinn er *hóflegur og skilyrtur* — kemur að mestu fram eftir fínþjálfun og sést skýrast í heild og undir álagi — ekki koma alhliða vélmennis sem má sleppa beint út í heiminn.

Rólegri en mikilvægari merkingin er aðferðafræðileg. Greinin er í raun mælistika: hún sýnir hversu mikið af gögnum þarf til að fullyrða áreiðanlega að ein vélmennastefna sé betri en önnur og gefur í skyn að töluverður hluti spennandi niðurstaðna sviðsins byggist á of litlu úrtaki. Sú leiðrétting er verðmætari en enn eitt líkanið.

Hvað sýnir þetta *ekki*?

- Þetta er **ekki alhliða vélmenni**. Ávinningurinn er sýndur fyrir tiltekna byggingu — diffusion-stefnur — sem eru fínþjálfaðar fyrir hvert verkefni, við stýrðar aðstæður og út frá fjarsýrðum sýnidæmum; ekki sjálfstætt vélmenni sem tekur við hvaða nýju verkefni sem er samkvæmt skipun.
- Það staðfestir **ekki zero-shot notkun**. Án fínþjálfunar vann stóra líkanið ekki grunngerðirnar stöðugt.
- Þetta eru **ekki gögn um „emergent stökk“**. Skölun bætti frammistöðu jafnt og þétt; enginn skarpur þröskuldur kom fram.
- Tölurnar eru **afstæðar og bundnar rannsóknarstofu**. Heildarárangur var vísvitandi stilltur nálægt ~50% svo auðveldara væri að greina samanburðarmun; þetta er ekki mæling á raunverulegum áreiðanleika og rannsóknin skoðar eina byggingu frá einni rannsóknarstofu.

- Hún skýrir **ekki hvers vegna** einstök verkefni takast eða mistakast, og nokkur verkefni þar sem stóra líkanið stóð sig verr eru birt en ekki útskýrð.
- Hún segir **ekkert um öryggi, sjálfstæði eða notkun** utan prófunarumhverfisins.

Hversu sterk eru gögnin?

Fyrir megin samanburðarniðurstöðurnar — að fínþjálfuð LBM vinni frá-grunni líkön í heild, þurfi margfalt minna gögn og séu seigari við dreifingarfærslu — eru gögnin sterk og óvenju vel stýrð: blint, slembiraðað, stór úrtök, formleg tölfræðipróf og gæðapróf á stigagjöf. Þetta er sjaldgæft tilvik þar sem aðferðafræðin er nógu sterk til að taka meginniðurstöðurnar nokkuð beint.

Heiðarlegu varnaglar eru þeir sem höfundarnir nefna sjálfir. Skekkjumörkin fanga tilviljun í *prófuninni* en **ekki tilviljun milli þjálfunarkeyrslna** — sama líkan þjálfað tvisvar getur gefið merkjanlega ólíka stefnu, og sá breytileiki er ekki í tölfræðinni. Hvert raunheimsverkefni hafði 50 prófanir, nóg fyrir meðalstór áhrif en mögulega of lítið fyrir lítil. Tungumálaskilyrðingin notaði fremur hóflegan kóðara, þannig að fullyrðingar um stærri „segðu vélmenninu bara hvað það á að gera“-kerfi kunna að verða aðrar. Og síðan er hreinskilin uppljóstrun um normaliserunarvillu sem fannst eftir á. Ekkert af þessu fellir meginniðurstöðuna, en þetta eru nákvæmlega atriðin sem greinin segir að sviðið hunsi oft.

Ein heimildanóta í sama anda: þessi útskýring byggir á forprenti höfundanna. Við gátum ekki sótt birtu tímaritsútgáfuna og höfum því ekki borið saman hvort breytingar urðu milli forprents og lokaútgáfu.

Hvers vegna skiptir þetta máli?

„Vélmennagrunnlíkan“ er hugtak sem býður upp á ofmat, og auðvelt er að ranglesa rannsóknina í báðar áttir — sem sigurhróp „þetta virkar!“ eða afskrift „þetta er bara hype“. Nákvæmara svarið er gagnlegra: forþjálfun á fjölbreyttum gögnum gefur raunverulegan, mælanlegan en hóflegan ávinning, sérstaklega minna gagnamagn á verkefni og meiri seiglu; og ávinningurinn eykst fyrirsjáanlega með skala.

Dýpri ástæðan er að greinin beinir strangleikanum að eigin fræðigrein. Með því að sýna að raunverulegar áhrif eru nógu lítil til að hverfa í slæmu mati, og að leiðinleg ákvörðun eins og gagnanormalisering getur skipt meira máli en snjöll ný bygging, setur hún fram rök fyrir því að framfarir í vélmennanámi þurfi mun traustari mælingar áður en þær eru trúverðugar. Grein sem eyðir trúverðugleika sínum í að gæta mörkanna milli niðurstöðu og óskar er að gera eitthvað sjaldgæfara og verðmætara en að toppa stigatöflu.

Í stuttu máli

Vísindamenn hjá Toyota Research Institute þjálfuðu „large behavior models“ — diffusion-byggðar vélmennastefnur forþjálfaðar á um 1.700 klukkustundum af fjölbreyttum hreyfigögnum — og prófuðu þær gegn eins-verkefnislíkönum frá grunni með óvenju ströngu ferli: blint, slembiraðað, stórt úrtak (≈1.800 raunheims- og yfir 47.000 hermunarprófanir) og formleg tölfræði. Eftir fínþjálfun fyrir hvert verkefni stóðu stóru líkönin sig betur í heild, þurftu um **3–5× minna verkefnissértækt gagnamagn** og voru **seigari þegar aðstæður breyttust**. Frammistaða jókst jafnt með meiri forþjálfun. En **án fínþjálfunar** unnu þau ekki eins-verkefnislíkön stöðugt; mörg áhrif voru svo lítil að stóru úrtökin þurfti til að sjá þau; og hversdagsleg gagnanormalisering skipti meira máli en byggingarbreytingar. Þetta er traustur en hófstilltur stuðningur við vélmennagrunnlíkön — **ekki** alhliða vélmenni, ekki zero-shot alhæfari og ekki „emergent stökk“ — ásamt beinni viðvörun um að stór hluti vélmennafræðinnar kunni að vera að mæla hávaða.

Raunhæf yfirferð

Það sem greinin sýnir: Með ströngu blindu og tölfræðilega nægu mati (≈1.800 raunheimur + 47.000+ hermunarkeyrslur) vinna fjölverkefna-forþjálfaðar og síðan fínþjálfaðar diffusion-stefnur (LBM) frá-grunni eins-verkefnisstefnur í heild, ná sambærilegri frammistöðu með ~3–5× minna verkefnissértæku gögnum og þola dreifingarfærslu betur; frammistaða vex slétt með forþjálfunargögnum.

Það sem er líklegt en ekki sannað: Að sami ávinningur flytjist yfir í mun stærri vision-language-action líkön; tungumálakóðari þeirra var lítill. Einnig að slétt skölun haldi áfram langt út fyrir gagnasviðið sem var prófað.

Það sem greinin sýnir ekki: Alhliða eða zero-shot vélmenni; neitt „emergent“ hæfileikastökk; skýringar á öllum einstökum verkefnabilunum; raunverulegan áreiðanleika í algildum tölum; eða neitt um öryggi og sjálfstæða notkun.

Helstu takmarkanir: Tölfræðin nær yfir prófunartilviljun en ekki breytileika milli þjálfunarkeyrslna; 50 raunheimspróf á verkefni geta misst lítil áhrif; ein bygging og ein rannsóknarstofa; hóflegur tungumálakóðari; normaliserunarvilla fannst eftir mat; og greiningin hér byggir á forprentinu, ekki staðfestri samanburðarskoðun við birtu útgáfuna.

Hversu mikið traust ætti almennur lesandi að bera til niðurstöðunnar? Mikið traust á því að fjölverkefnaforþjálfun + fínþjálfun gefi raunverulegan, hóflegan ávinning, sérstaklega í gagnanýtni og seiglu, og að hann hafi verið mældur óvenju vandlega. Mikið traust á því að þetta sé **ekki** alhliða eða zero-shot vélmenni og **ekki** emergent stökk. Miðlungs traust á því hversu langt ávinningurinn skalar í stærri kerfi. Og viðvörun höfundanna um að of lítil úrtök í vélmennafræði geti verið að mæla hávaða er þess virði að taka alvarlega.

Heimildir

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/