The Clean Paper

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

Fungerer store «grunnmodeller» for roboter faktisk bedre? Et grundig svar — et forsiktig ja, og de fleste studier klarer ikke å avgjøre det

Toyota Research Institute trente «store atferdsmodeller» — robotpolicyer forhåndstrent på ~1,700 timer med varierte manipulasjonsdata — og testet dem mot policyer for én oppgave som var trent fra bunnen av, med uvanlig grundighet: blindede, randomiserte forsøk med store utvalg (~1,800 i den virkelige verden, 47,000+ i simulering) og reell statistikk. Etter finjustering for hver oppgave gjorde de store modellene det bedre i gjennomsnitt, trengte omtrent 3–5× mindre oppgavespesifikke data og var mer robuste når forholdene endret seg; ytelsen steg jevnt med mer data til forhåndstreningen. Men uten finjustering slo de ikke konsekvent modeller for én oppgave, flere effekter var så små at de store utvalgene måtte til for at de i det hele tatt skulle bli synlige, og et hverdagslig valg om datanormalisering betydde mer enn arkitekturen. Dette er avmålt støtte til retningen med grunnmodeller for roboter — ikke en robot for generell bruk, ikke en zero-shot-generalist og ikke et «emergent sprang» — samt en tydelig advarsel om at mye av robotikken kanskje måler støy.

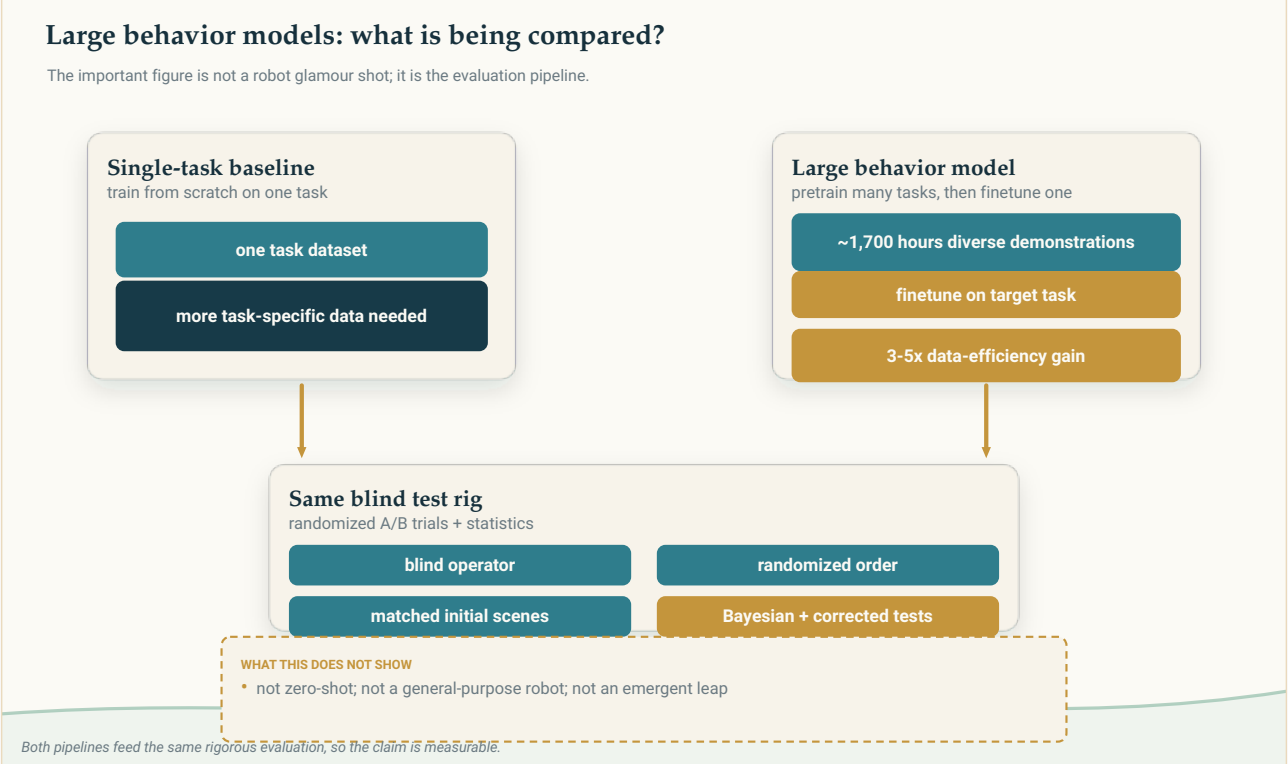
Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

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

En robotikkartikkel som egentlig handler om ærlighet

Robotikken er midt i et kappløp om «grunnmodeller». Ideen, lånt fra språk- og bilde-KI, er fristende: I stedet for å trene en robot til én oppgave om gangen trener man én **«stor atferdsmodell» (LBM)** på en enorm og variert mengde demonstrasjoner og får et system med bred kapasitet som raskt kan tilpasses. Entusiasmen — og investeringene — er enorm. Overskriften skriver seg selv: *robothjerner for generell bruk er her.*

Denne artikkelen fra Toyota Research Institute er interessant nettopp fordi den nekter å skrive den overskriften. Det egentlige bidraget er ikke en mer spektakulær robot. Det er en grundig undersøkelse av et tilsynelatende enkelt spørsmål — *fungerer tilnærmingen med store modeller faktisk bedre, og hvordan skulle vi i det hele tatt vite det?* — besvart med en statistisk grundighet som ifølge forfatterne selv er uvanlig på feltet. Resultatet er et genuint nyttig «ja, men» og en advarsel om at mye av robotikken kanskje måler støy.



Begge tilnærmingene går inn i den samme blindede, randomiserte evalueringen. Artikkelens påstand er ikke at grunnmodeller for roboter er zero-shot-generalister, men at forhåndstrening kan forbedre dataeffektivitet og robusthet når det måles grundig.

Hva forfatterne gjorde

De bygget LBM-er i en bestemt, konkret forstand: **diffusjonsbaserte visuomotoriske policyer** (en Diffusion Transformer som leser kamerabilder, en kort språkinstruksjon og robotens egne leddposisjoner og sender ut korte serier med motorkommandoer ved 10 Hz). Disse ble **forhåndstrent på omtrent 1,700 timer** med robotdemonstrasjoner — over 500 forskjellige oppgaver samlet inn internt, i tillegg til offentlige datasett — og deretter **finjustert** for enkeltoppgaver. Gjennomgående sammenlignes de med en **policy for én oppgave, trent fra bunnen av** på data fra akkurat denne oppgaven.

Kjernen i artikkelen er imidlertid *evalueringen*, som forfatterne behandler som hovedresultatet. For å unngå å lure seg selv brukte de:

- **Blindede, randomiserte A/B-tester** i den virkelige verden — personen som kjørte roboten, visste ikke hvilken policy som ble testet, og rekkefølgen var randomisert.
- **Kontrollerte, repeterbare startforhold** — operatørene tilpasset scenen til et bildeoverlegg før hvert forsøk.
- **Et stort antall forsøk** — 50 kjøringer i den virkelige verden per oppgave, per policy og per betingelse; 200 per oppgave i simulering. Totalt: omtrent **1,800 blindede kjøringer i den virkelige verden og mer enn 47,000 simulerte kjøringer**.
- **Korrekt statistikk** — bayesianske estimater av sannsynligheten for suksess og parvise hypotesetester med korreksjoner for multiple sammenligninger, i stedet for å vurdere overlappende feilmarginer på øyemål. De gjennomførte til og med en kvalitetssikringsrunde for en fjerdedel av forsøkene som mennesker hadde vurdert, for å måle vurderingsfeil.

[video pending]

Sammenligning side om side av modeller som dekker et frokostbord: (venstre) referansemodellen for én oppgave og (høyre) LBM. Begge videoene spilles av i 1x hastighet. Dette er én evaluert oppgave, ikke et bevis på autonomi for generell bruk.

Dette apparatet er selve poenget. Hele artikkelen argumenterer for at man uten det ikke kan skille en reell forbedring fra flaks.

Hva de fant

Finjusterte store modeller slår modeller for én oppgave som er trent fra bunnen av — i gjennomsnitt. Samlet på tvers av oppgaver presterte en LBM som var forhåndstrent og deretter finjustert for en oppgave, pålitelig bedre enn en policy som var trent fra bunnen av for den samme oppgaven, både i simulering og i den virkelige verden, og forskjellen var statistisk signifikant. På enkeltoppgaver var den finjusterte LBM-en statistisk like god som eller bedre enn modellen som var trent fra bunnen av, i nesten alle tilfeller (3/3 oppgaver i den virkelige verden, 15/16 simulerte oppgaver).

Den største og tydeligste gevinsten er dataeffektivitet. En finjustert LBM nådde samme ytelse som en modell trent fra bunnen av med omtrent **3–5× mindre oppgavespesifikke data**. I én oppgave i

den virkelige verden (å dekke et frokostbord) slo en LBM som var finjustert på bare **15%** av demonstrasjonene, en policy som var trent fra bunnen av på **100%** av dem.

Forhåndstrening hjelper mest når forholdene endrer seg. Da testmiljøet med vilje ble forskjøvet bort fra treningsforholdene («distribusjonsskift»), *økte* den finjusterte LBM-ens forsprang. I ett simuleringssett slo den statistisk modellen som var trent fra bunnen av, i 3 av 16 oppgaver under normale forhold, men i 10 av 16 ved distribusjonsskift. Siden virkelige driftsmiljøer alltid glir bort fra treningsforholdene, kan denne robustheten hevdes å være det praktisk viktigste funnet.

Mer data til forhåndstrening hjalp, på en jevn måte. Ytelsen steg jevnt etter hvert som de la til data til forhåndstreningen — uten **noe plutselig hopp eller «emergent» sprang** ved skalaene som ble testet. Nyttig, forutsigbart og udramatisk.

Men fortellingen om en generalist uten finjustering holdt ikke. En forhåndstrent LBM brukt *zero-shot* — uten oppgavespesifikk finjustering — slo *ikke* konsekvent policyer for én oppgave. Ett enkelt nettverk kunne gjøre mange oppgaver samtidig, men drømmen om å «bare gi den en instruksjon» fikk ikke støtte her; forfatterne tilskriver dette delvis den begrensede språkenkoderens sårbarhet.

Og gevinstene var små nok til å være lette å overse — eller å fremstille på feilaktig grunnlag. Mange av effektene ble først synlige med større utvalg enn vanlig og grundige tester. Forfatterne sier rett ut at det, gitt størrelsen på effektene og støyen, **er en betydelig risiko for at mange robotikkartikler måler statistisk støy.** De fant også at et hverdagslig valg — hvordan dataene *normaliseres* — påvirket resultatene mer enn arkitekturendringer, og at en normaliserings**feil** i forhåndstreningen først ble oppdaget *etter* at evalueringene var fullført.

Hva dette sannsynligvis betyr

Den forsvarlige tolkningen: Forhåndstrening i stor skala på varierte robotdata er en reell og verdifull bestanddel — den gjør at man trenger mindre data per ny oppgave, og den gjør policyene mer robuste når verden ikke samsvarer med treningen. Det gir genuin støtte til retningen feltet satser på. Men gevinstene er *moderate og betingede* (de viser seg hovedsakelig etter finjustering og er tydeligst samlet og under belastning), ikke et tegn på at en ferdig, generell robot som bare kan tas i bruk, har ankommet.

Den mindre iøynefallende og viktigere betydningen er metodologisk. Artikkelen fungerer i praksis som en målestokk: Den viser hvor mye dokumentasjon som faktisk kreves for å fremsette en troverdig påstand om en robotpolicy, og antyder at mye av den publiserte begeistringen hviler på for lite. Det er en korrigering feltet trenger mer enn enda en modell.

Hva dette ikke beviser

- Det er **ikke en robot for generell bruk.** Gevinstene er vist for en bestemt arkitektur (diffusjonspolicyer), finjustert for hver oppgave, i kontrollerte omgivelser og basert på fjernstyrte demon-

strasjoner — ikke for en autonom robot som utfører vilkårlige nye jobber på kommando.
- Det bekrefter **ikke** bruk med **zero-shot**. Uten finjustering slo den store modellen ikke konsekvent referansemodellene for én oppgave.
- Det er **ikke** bevis for et **«emergent sprang».** Oppskalering ga jevne forbedringer; det finnes ingen diskontinuitet her som støtter fortellinger om «og så ble den plutselig kompetent».
- Tallene er **relative og bundet til laboratoriet.** De absolutte suksessratene ble med vilje justert mot ~50% for å gjøre sammenligningene følsomme; de måler ikke pålitelighet i den virkelige verden, og arbeidet gjelder én arkitektur fra ett laboratorium.
- Det avgjør **ikke hvorfor** en policy lykkes eller mislykkes, og flere konkrete oppgaver der den store modellen gjorde det *dårligere,* rapporteres uten å bli forklart.
- Det sier **ingenting om sikkerhet, autonomi eller utrulling** utenfor evalueringsriggen.

Hvor sterke er bevisene?

For de sentrale sammenlignende påstandene — at finjusterte LBM-er samlet slår referansemodeller trent fra bunnen av, trenger flere ganger mindre data og er mer robuste ved distribusjonsskift — er bevisene sterke *og* uvanlig godt kontrollerte: blindet, randomisert, med store utvalg, statistisk testet og med en kvalitetssikringsrunde av vurderingen. Dette er det sjeldne tilfellet der metodikken er solid nok til at hovedkonklusjonene kan tas omtrent som de står.

De ærlige forbeholdene er slike forfatterne selv tar opp. Feilmarginene deres fanger tilfeldigheten i *evalueringen,* men **ikke** tilfeldigheten i *treningen* — trener man den samme modellen to ganger, kan man få policyer som er merkbart forskjellige, og denne variasjonen inngår ikke i statistikken. Oppgavene i den virkelige verden hadde 50 forsøk hver, nok til å oppdage middels store effekter, men med risiko for å overse små. Språkstyringen brukte en beskjeden enkoder, så påstander om å «bare fortelle roboten hva den skal gjøre» kan slå annerledes ut for større systemer. I tillegg kommer den oppriktige opplysningen om en normaliseringsfeil som ble oppdaget i etterkant. Ingenting av dette undergraver hovedfunnene, men det er nettopp slike ting artikkelen hevder at feltet vanligvis feier til side.

En merknad om kilden, i samme ånd: Denne forklaringen er basert på forfatternes preprint. Vi klarte ikke å hente den tidsskriftpubliserte versjonen og har derfor ikke kontrollert om noe ble endret mellom preprinten og den publiserte teksten.

Hvorfor det er viktig

«Grunnmodeller for roboter» er et uttrykk som innbyr til overdrevne påstander, og en studie som denne er lett å feiltolke i begge retninger — som et triumferende «*det fungerer!*» eller et avvisende «*det er overhypet*». Den presise konklusjonen er mer nyttig enn begge: Forhåndstrening på varierte data gir reelle, målbare, men moderate fordeler — først og fremst mindre data per oppgave og større robusthet — og utviklingen forbedres forutsigbart med skalaen.

Den dypere grunnen til at artikkelen betyr noe, er at den retter sin grundighet mot sitt eget felt. Ved å vise at de reelle effektene er små nok til å forsvinne i en slurvete evaluering, og at et kjedelig valg som datanormalisering kan veie tyngre enn en smart ny arkitektur, argumenterer den for at mye av fremgangen innen robotlæring trenger mer robuste målinger før den kan tros. En artikkel som bruker troverdigheten sin til å vokte forskjellen mellom et resultat og et ønske, gjør noe sjeldnere og mer verdifullt enn å toppe en resultatliste.

Kort oppsummert

Forskere ved Toyota Research Institute trente «store atferdsmodeller» — diffusjonsbaserte robotpolicyer forhåndstrent på ~1,700 timer med varierte manipulasjonsdata — og testet dem mot policyer for én oppgave som var trent fra bunnen av, med en uvanlig grundig protokoll: blindet, randomisert og med store utvalg (≈1,800 forsøk i den virkelige verden og 47,000+ simulerte forsøk), med reell statistikk. Etter finjustering for hver oppgave gjorde de store modellene det pålitelig bedre samlet, trengte omtrent **3–5× mindre oppgavespesifikke data** og var **mer robuste når forholdene endret seg**, samtidig som ytelsen steg jevnt etter hvert som mengden data til forhåndstreningen økte. Men brukt **uten finjustering** slo de *ikke* konsekvent modeller for én oppgave, flere effekter var så små at bare de store utvalgene avdekket dem, og et hverdagslig valg om datanormalisering betydde mer enn arkitekturen. Dette er solid, avmålt støtte til retningen med grunnmodeller for roboter — **ikke** en robot for generell bruk, ikke en zero-shot-generalist og ikke et «emergent sprang» — samt en tydelig advarsel om at mye av robotikken kanskje måler støy.

Nøktern vurdering

Hva artikkelen viser: Med en grundig, blindet evaluering med tilstrekkelig statistisk styrke (≈1,800 kjøringer i den virkelige verden + 47,000+ simulerte kjøringer) presterer diffusjonspolicyer (LBM-er) som er forhåndstrent på flere oppgaver og deretter finjustert, samlet bedre enn policyer for én oppgave som er trent fra bunnen av, når tilsvarende ytelse med ~3–5× mindre oppgavespesifikke data og er mer robuste ved distribusjonsskift; ytelsen skalerer jevnt med mengden data til forhåndstreningen.

Hva som er plausibelt, men ikke bevist: At disse fordelene overføres til mye større modeller for syn, språk og handling (språkenkoderen deres var liten); at den jevne skaleringen fortsetter utover det testede dataområdet.

Hva den ikke viser: En generell robot eller zero-shot-robot (ingen finjustering → ingen konsekvent fordel); noe «emergent» sprang i evne; forklaringer på konkrete feil på oppgavenivå; pålitelighet i den virkelige verden i absolutte tall (suksessratene ble justert til nær 50% for følsomhet); noe om sikkerhet eller autonom utrulling.

Viktigste begrensninger: Statistikken fanger tilfeldighet i evalueringen, men ikke mellom treningskjøringer; 50 forsøk i den virkelige verden per oppgave kan overse små effekter; én arkitektur

og ett laboratorium; beskjeden språkenkoder; en datanormaliseringsfeil ble oppdaget etter evalueringene; analysen er basert på preprinten (den publiserte versjonen er ikke kontrollert).

Hvor stor tillit bør en vanlig leser ha? Høy tillit til at forhåndstrening på flere oppgaver pluss finjustering gir reelle, moderate fordeler — særlig dataeffektivitet og robusthet — og at disse ble målt uvanlig grundig. Høy tillit til at dette **ikke** er en generell robot eller zero-shot-robot og **ikke** et emergent sprang. Middels tillit til hvor langt gevinstene skalerer til større modeller. Og dette er verdt å ta alvorlig: forfatternes egen advarsel om at effektene på feltet er små nok til at studier med for lav statistisk styrke kan rapportere støy. En passende holdning er avmålt optimisme om tilnærmingen og sunn skepsis til resultater innen robot-KI som mangler denne typen statistisk forankring.

Kilder

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/