The Clean Paper

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

Virker store "grundmodeller" til robotter faktisk bedre? Et grundigt svar — et beskedent ja, og de fleste studier kan ikke afgøre det

Toyota Research Institute trænede "store adfærdsmodeller" — robotpolitikker fortrænet på ~1,700 timer med varierede manipulationsdata — og testede dem mod politikker til én opgave, som var trænet fra bunden, med usædvanlig omhu: blindede, randomiserede forsøg med store stikprøver (~1,800 i den virkelige verden, 47,000+ i simulering) og reel statistik. Efter finjustering til hver opgave klarede de store modeller sig bedre i gennemsnit, behøvede omtrent 3–5× mindre opgavespecifikke data og var mere robuste, når betingelserne ændrede sig; ydeevnen steg jævnt med flere fortræningsdata. Men uden finjustering slog de ikke konsekvent modeller til én opgave, flere effekter var så små, at de store stikprøver var nødvendige for overhovedet at se dem, og et hverdagsagtigt valg om datanormalisering betød mere end arkitekturen. Det er afmålt støtte til retningen med grundmodeller til robotter — ikke en robot til generelle formål, ikke en zero-shot-generalist og ikke et "emergent spring" — samt en skarp advarsel om, at en stor del af robotforskningen måske måler støj.

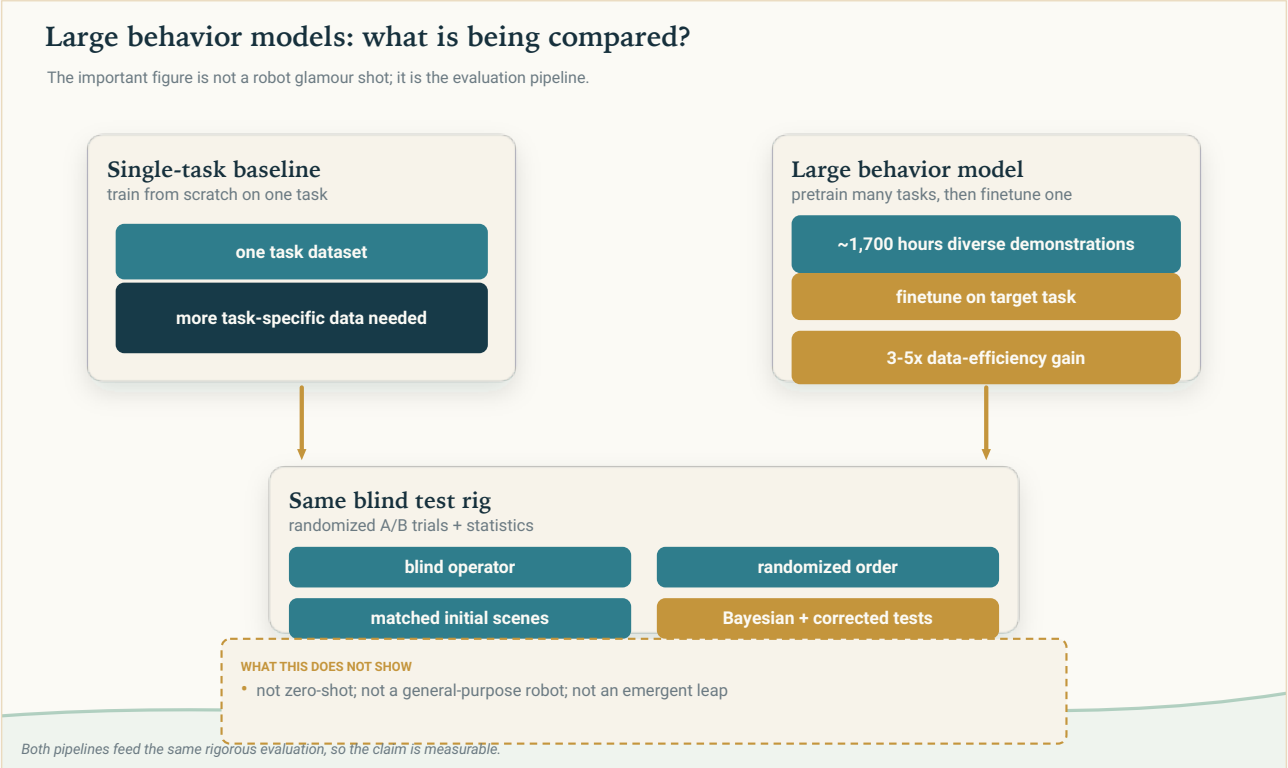
Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso · AI-assisted, human-reviewed

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 robotartikel, hvis egentlige emne er ærlighed

Robotteknologien befinder sig midt i et kapløb om "grundmodeller". Ideen, der er lånt fra sprog- og billed-AI, er tillokkende: I stedet for at træne en robot til én opgave ad gangen træner man én **"stor adfærdsmodel" (LBM)** på en enorm og varieret mængde demonstrationer og får et system med brede evner, som hurtigt kan tilpasses. Begejstringen — og investeringerne — er enorm. Overskriften skriver sig selv: *robothjerner til generelle formål er her.*

Denne artikel fra Toyota Research Institute er interessant, netop fordi den nægter at skrive den overskrift. Dens egentlige bidrag er ikke en mere imponerende robot. Det er en grundig undersøgelse af et bedragerisk enkelt spørgsmål — *virker tilgangen med store modeller faktisk bedre, og hvordan skulle vi overhovedet vide det?* — besvaret med en statistisk omhu, der ifølge forfatterne selv er usædvanlig på området. Resultatet er et oprigtigt nyttigt "ja, men" og en advarsel om, at en stor del af robotforskningen måske måler støj.



Begge tilgange føres ind i den samme blindede, randomiserede evaluering. Artiklens påstand er ikke, at grundmodeller til robotter er zero-shot-generalister, men at fortræning kan forbedre dataeffektivitet og robusthed, når der måles omhyggeligt.

Hvad forfatterne gjorde

De byggede LBM'er i en bestemt, konkret forstand: **diffusionsbaserede visuomotoriske politikker** (en Diffusion Transformer, der læser kamerabilleder, en kort sproglig instruktion og robottens egne ledpositioner og udsender korte serier af motorkommandoer ved 10 Hz). Disse blev **fortrænet på omtrent 1,700 timer** med robotdemonstrationer — over 500 forskellige opgaver indsamlet internt samt offentlige datasæt — og derefter **finjusteret** til enkelte opgaver. Sammenligningen er hele vejen igennem med en **politik til én opgave, trænet fra bunden** på data fra netop den opgave.

Artiklens kerne er imidlertid *evalueringen*, som forfatterne behandler som hovedresultatet. For at undgå at narre sig selv brugte de:

- **Blindede, randomiserede A/B-test** i den virkelige verden — den person, der betjente robotten, vidste ikke, hvilken politik der blev testet, og rækkefølgen var randomiseret.
- **Kontrollerede, gentagelige startbetingelser** — operatørerne tilpassede scenen til en billedoverlejring før hvert forsøg.
- **Et stort antal forsøg** — 50 kørsler i den virkelige verden per opgave, per politik og per betingelse; 200 per opgave i simulering. I alt: omkring **1,800 blindede kørsler i den virkelige verden og mere end 47,000 simulerede kørsler**.
- **Korrekt statistik** — bayesianske estimater af sandsynligheden for succes og parvise hypotesetest med korrektion for multiple sammenligninger frem for at vurdere overlappende fejlsøjler på øjemål. De gennemførte endda en kvalitetssikringsrunde på en fjerdedel af de forsøg, som mennesker havde bedømt, for at måle bedømmelsesfejl.

[video pending]

Sammenligning side om side af modeller, der dækker et morgenmadsbord: (venstre) baseline til én opgave og (højre) LBM. Begge videoer afspilles ved 1x hastighed. Dette er én evalueret opgave, ikke et bevis på autonomi til generelle formål.

Dette apparat er selve pointen. Hele artiklen er et argument for, at man uden det ikke kan skelne en reel forbedring fra held.

Hvad de fandt

Finjusterede store modeller slår modeller til én opgave, som er trænet fra bunden — i gennemsnit. Samlet på tværs af opgaver klarede en LBM, der var fortrænet og derefter finjusteret til en opgave, sig pålideligt bedre end en politik, der var trænet fra bunden til samme opgave, både i simulering og i den virkelige verden, og forskellen var statistisk signifikant. På enkeltopgaver var den finjusterede LBM statistisk set lige så god som eller bedre end modellen trænet fra bunden i næsten alle tilfælde (3/3 opgaver i den virkelige verden, 15/16 simulerede opgaver).

Den største og tydeligste gevinst er dataeffektivitet. En finjusteret LBM nåede samme ydeevne som en model trænet fra bunden med omtrent **3–5× mindre opgavespecifikke data**. I én opgave i den

virkelige verden (at dække et morgenmadsbord) slog en LBM, der var finjusteret på blot **15%** af demonstrationerne, en politik trænet fra bunden på **100%** af dem.

Fortræning hjælper mest, når betingelserne ændrer sig. Da testmiljøet bevidst blev forskudt væk fra træningsbetingelserne (”distributionsskift”), *voksede* den finjusterede LBM’s forspring. I ét sæt simuleringer slog den statistisk modellen trænet fra bunden på 3 af 16 opgaver under normale betingelser, men på 10 af 16 ved distributionsskift. Eftersom virkelige driftsmiljøer altid glider væk fra træningsbetingelserne, kan denne robusthed hævdes at være det praktisk vigtigste resultat.

Flere fortræningsdata hjalp, på en jævn måde. Ydeevnen steg støt, efterhånden som de tilføjede fortræningsdata — uden **noget pludseligt spring eller ”emergent” gennembrud** ved de afprøvede skalaer. Nyttigt, forudsigeligt og udramatisk.

Men fortællingen om en generalist uden finjustering holdt ikke. En fortrænet LBM anvendt *zero-shot* — uden opgavespecifik finjustering — slog *ikke* konsekvent politikker til én opgave. Ét enkelt netværk kunne udføre mange opgaver på én gang, men drømmen om ”bare at bede den om det” blev ikke underbygget her; forfatterne tilskriver dette delvist skrøbeligheden i deres begrænsede sprogindkoder.

Og gevinsterne var små nok til let at blive overset — eller fremstillet på falsk grundlag. Mange af effekterne blev først synlige med større stikprøver end normalt og omhyggelige test. Forfatterne siger direkte, at der i betragtning af effekternes størrelse og støjen **er en betydelig risiko for, at mange robotartikler måler statistisk støj.** De fandt også, at et hverdagsagtigt valg — hvordan dataene *normaliseres* — påvirkede resultaterne mere end arkitekturændringer, og at en normaliserings**fejl** i fortræningen først viste sig, *efter* at evalueringerne var afsluttet.

Hvad dette sandsynligvis betyder

Den forsvarlige læsning: Fortræning i stor skala på varierede robotdata er en reel og værdifuld bestanddel — den mindsker behovet for data til hver ny opgave og gør politikkerne mere robuste, når verden ikke svarer til træningen. Det giver oprigtig støtte til den retning, som feltet satser på. Men gevinsterne er *moderate og betingede* (de viser sig hovedsageligt efter finjustering og er tydeligst samlet og under belastning), ikke ankomsten af en brugsklar, generel robot.

Den mere afdæmpede og vigtigere betydning er metodologisk. Artiklen fungerer i praksis som en målestok: Den viser, hvor meget dokumentation der faktisk skal til for at fremsætte en troværdig påstand om en robotpolitik, og antyder, at meget af den publicerede begejstring hviler på for lidt. Det er en korrektion, som feltet har mere brug for end endnu en model.

Hvad dette *ikke* beviser

- Det er **ikke en robot til generelle formål.** Gevinsterne er påvist for en bestemt arkitektur (diffusionspolitikker), finjusteret til hver opgave, i kontrollerede omgivelser og ud fra teleopererede

demonstrationer — ikke for en autonom robot, der udfører vilkårlige nye opgaver på kommando.
- Det validerer **ikke** brug med **zero-shot**. Uden finjustering slog den store model ikke konsekvent baselines til én opgave.
- Det er **ikke** evidens for et **"emergent spring".** Opskalering forbedrede resultaterne jævnt; her er ingen diskontinuitet, der understøtter fortællinger om, at "og så blev den pludselig i stand til det".
- Tallene er **relative og bundet til laboratoriet.** De absolutte succesrater blev bevidst justeret mod ~50% for at gøre sammenligningerne følsomme; de er ikke et mål for pålidelighed i den virkelige verden, og arbejdet omfatter én arkitektur fra ét laboratorium.
- Det afgør **ikke hvorfor** en politik lykkes eller fejler, og flere konkrete opgaver, hvor den store model klarede sig *dårligere*, rapporteres uden at blive forklaret.
- Det siger **intet om sikkerhed, autonomi eller ibrugtagning** uden for evalueringsriggen.

Hvor stærk er evidensen?

For de centrale sammenlignende påstande — at finjusterede LBM'er samlet slår baselines trænet fra bunden, behøver flere gange mindre data og er mere robuste ved distributionsskift — er evidensen stærk *og* usædvanligt velkontrolleret: blindet, randomiseret, med store stikprøver, statistisk testet og med en kvalitetssikringsrunde af bedømmelserne. Dette er det sjældne tilfælde, hvor metoden er solid nok til, at hovedkonklusionerne kan tages næsten for pålydende.

De ærlige forbehold er dem, forfatterne selv fremhæver. Deres fejlsøjler indfanger tilfældigheden i *evalueringen*, men **ikke** tilfældigheden i *træningen* — træner man den samme model to gange, kan man få politikker, der adskiller sig mærkbart, og den variation indgår ikke i statistikken. Opgaverne i den virkelige verden havde 50 forsøg hver, nok til at opdage mellemstore effekter, men med risiko for at overse små. Sprogstyringen brugte en beskeden indkoder, så påstande om "bare at fortælle robotten, hvad den skal gøre" kan falde anderledes ud for større systemer. Dertil kommer den åbne oplysning om en normaliseringsfejl, som blev fundet efterfølgende. Intet af dette vælter hovedresultaterne, men det er netop den slags, artiklen hævder, at feltet normalt fejer til side.

En bemærkning om kilden i samme ånd: Denne forklaring er baseret på forfatternes preprint. Vi kunne ikke hente den tidsskriftspublicerede version og har derfor ikke kontrolleret, om noget blev ændret mellem preprinten og den publicerede tekst.

Hvorfor det er vigtigt

"Grundmodeller til robotter" er et udtryk, der er skabt til overdrevne påstande, og en undersøgelse som denne er let at fejllæse i begge retninger — som et triumferende *"det virker!"* eller et afvisende *"det er overhypet"*. Den præcise konklusion er mere nyttig end begge: Fortræning på varierede data giver reelle, målbare, men moderate fordele — først og fremmest mindre data per opgave og større robusthed — og udviklingen forbedres forudsigeligt med skalaen.

Den dybere grund til, at artiklen er vigtig, er, at den retter sin omhu mod sit eget felt. Ved at vise,

at de reelle effekter er små nok til at forsvinde i en sjusket evaluering, og at et kedeligt valg som datanormalisering kan veje tungere end en snedig ny arkitektur, argumenterer den for, at en stor del af fremskridtene inden for robotlæring behøver mere robuste målinger, før man kan tro på dem. En artikel, der bruger sin troværdighed på at vogte over forskellen mellem et resultat og et ønske, gør noget sjældnere og mere værdifuldt end at toppe en rangliste.

Kort og klart

Forskere ved Toyota Research Institute trænede "store adfærdsmodeller" — diffusionsbaserede robotpolitikker fortrænet på ~1,700 timer med varierede manipulationsdata — og testede dem mod politikker til én opgave, der var trænet fra bunden, med en usædvanligt stringent protokol: blindet, randomiseret og med store stikprøver (≈1,800 forsøg i den virkelige verden og 47,000+ simulerede forsøg), med reel statistik. Efter finjustering til hver opgave klarede de store modeller sig pålideligt bedre samlet, behøvede omtrent **3–5× mindre opgavespecifikke data** og var **mere robuste, når betingelserne ændrede sig**, mens ydeevnen steg jævnt, i takt med at mængden af fortræningsdata voksede. Men anvendt **uden finjustering** slog de *ikke* konsekvent modeller til én opgave, flere effekter var så små, at kun de store stikprøver afslørede dem, og et hverdagsagtigt valg om datanormalisering betød mere end arkitekturen. Det er solid, afmålt støtte til retningen med grundmodeller til robotter — **ikke** en robot til generelle formål, ikke en zero-shot-generalist og ikke et "emergent spring" — samt en skarp advarsel om, at en stor del af robotforskningen måske måler støj.

Tjek uden omsvøb

Hvad artiklen viser: Med en stringent, blindet evaluering med tilstrækkelig statistisk styrke (≈1,800 kørsler i den virkelige verden + 47,000+ simulerede kørsler) klarer diffusionspolitikker (LBM'er), som er fortrænet på flere opgaver og derefter finjusteret, sig samlet bedre end politikker til én opgave, der er trænet fra bunden, opnår tilsvarende ydeevne med ~3–5× mindre opgavespecifikke data og er mere robuste ved distributionsskift; ydeevnen skalerer jævnt med mængden af fortræningsdata.

Hvad der er plausibelt, men ikke bevist: At disse fordele kan overføres til langt større modeller for syn, sprog og handling (deres sprogindkoder var lille); at den jævne skalering fortsætter ud over det testede dataområde.

Hvad den ikke viser: En generel robot eller zero-shot-robot (ingen finjustering → ingen konsekvent fordel); noget "emergent" spring i evner; forklaringer på konkrete fejl på opgaveniveau; pålidelighed i den virkelige verden i absolutte tal (succesraterne blev justeret til nær 50% for følsomhed); noget om sikkerhed eller autonom ibrugtagning.

Vigtigste begrænsninger: Statistikken indfanger tilfældighed i evalueringen, men ikke mellem træningskørsler; 50 forsøg i den virkelige verden per opgave kan overse små effekter; én arkitektur og ét laboratorium; beskeden sprogindkoder; en datanormaliseringsfejl blev fundet efter evalueringerne; analysen er baseret på preprinten (den publicerede version er ikke kontrolleret).

Hvor stor tillid bør en almindelig læser have? Høj tillid til, at fortræning på flere opgaver plus finjustering giver reelle, moderate fordele — især dataeffektivitet og robusthed — og at disse blev målt usædvanligt omhyggeligt. Høj tillid til, at dette **ikke** er en generel robot eller zero-shot-robot og **ikke** et emergent spring. Middel tillid til, hvor langt gevinsterne skalerer til større modeller. Og dette er værd at tage alvorligt: forfatternes egen advarsel om, at feltets effekter er små nok til, at undersøgelser med utilstrækkelig statistisk styrke kan rapportere støj. Den passende holdning er afmålt optimisme om tilgangen og sund skepsis over for resultater inden for robot-AI, som mangler denne form for statistisk underbygning.

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/