The Clean Paper

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

Shin manyan robot “foundation models” suna aiki fiye da gaske? Amsa mai hankali — eh, kaɗan, kuma yawancin studies ba za su iya gane bambancin ba

Toyota Research Institute ta horar da “large behavior models” — robot policies da aka pretrain a kan ~awa 1,700 na diverse manipulation data — sannan ta gwada su da from-scratch single-task policies da unusual rigour: blind, randomized, large-sample trials (~1,800 real-world, 47,000+ simulation) tare da real statistics. Bayan per-task finetuning, big models sun fi kyau a matsakaici, sun bukaci kusan 3–5× karancin task-specific data, kuma sun fi robust lokacin da conditions suka canza; performance ya karu smoothly da karin pretraining data. Amma ba tare da finetuning ba ba su consistently beat single-task models ba, effects da dama sun yi kankanta har sai da manyan samples suka bayyana su, kuma data-normalisation choice na yau da kullum ya fi architecture tasiri. Wannan measured support ne ga robot-foundation-model direction — ba general-purpose robot ba, ba zero-shot generalist ba, ba “emergent leap” ba — tare da gargadi cewa wani ɓangare mai yawa na robotics na iya kasancewa yana auna noise.

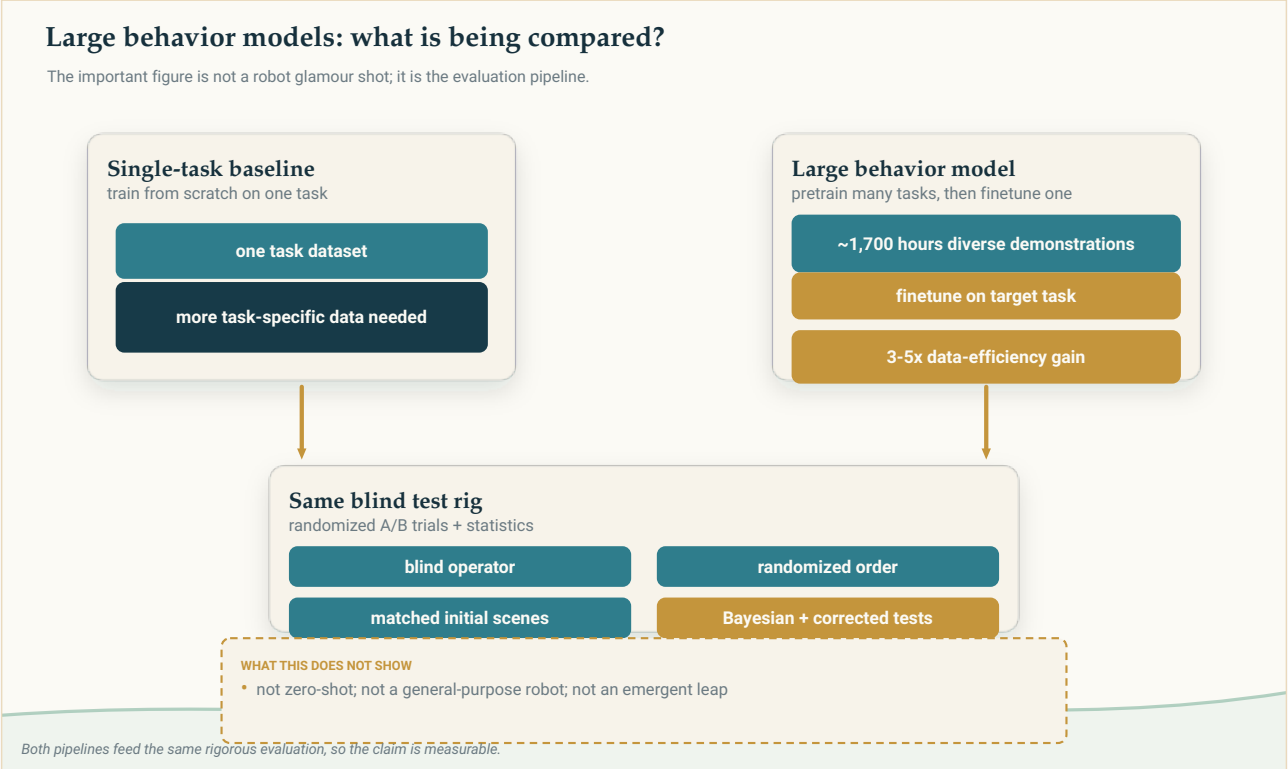
Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

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

Takardar robot da ainihin batunta shi ne gaskiya wajen au-nawa

Robotics tana tsakiyar wani yunƙurin gina “foundation models.” Ra’ayin, wanda aka aro daga language da image AI, yana da jan hankali: maimakon horar da robot kan aiki guda-guda, a horar da babban **large behavior model (LBM)** a kan tarin demonstrations masu yawa da bambance-bambance, sannan a samu system da zai iya ayyuka da yawa kuma ya saba da sabo da sauri. Sha’awar fannin da kuɗin da ake zuba suna da yawa. Kanun labarin da ya fi sauƙi shi ne: *general-purpose robot brains sun zo.*

Abin da ya sa wannan takarda daga Toyota Research Institute take da ban sha’awa shi ne ta ƙi rubuta wannan kanun labari. Babban gudummawarta ba robot mafi armashi ba ce. Ta ɗauki wata tambaya mai sauƙi a gani amma mai wahala wajen amsawa: **shin hanyar manyan models tana aiki fiye da yadda ake horar da task ɗaya, kuma ta yaya za mu tabbatar?** Marubutan sun amsa da matakin kula da kididdiga da, a cewarsu, ba kasafai ake gani a fannin ba. Sakamakon “eh, amma…” ne mai amfani, tare da gargadi cewa wani ɓangare na robotics na iya zama yana auna hayaniya ta kididdiga.



Hanyoyin biyu suna shiga blind, randomized evaluation iri ɗaya. Ikirarin takardar ba cewa robot foundation models zero-shot generalists ne ba; shi ne cewa pretraining na iya inganta amfani da bayanai da robustness idan an auna shi da kyau.

Abin da marubutan suka yi

Sun gina LBMs da ma'ana takamaimai: **diffusion-based visuomotor policies** — Diffusion Transformer da ke karanta hotunan cameras, gajeriyar umarnin harshe da joint positions na robot, sannan ya fitar da gajerun jerin motor commands a 10 Hz. An yi **pretraining da kusan awa 1,700** na robot demonstrations — sama da tasks 500 da aka tattara a cikin gida tare da public datasets — sannan a yi **finetuning** ga kowane task. Kwatantawar da aka yi a duk takardar ita ce da **single-task policy da aka horar daga farko** da bayanan wannan task kaɗai.

Amma zuciyar takardar ita ce **evaluation**, wanda marubutan suke ɗauka a matsayin babban sakamakon. Domin kada su yaudari kansu sun yi amfani da:

- **Blind, randomized A/B testing** a zahiri — mutumin da ke gudanar da robot bai san wace policy ake gwadawa ba, kuma an bazu order ɗin gwaje-gwajen.
- **Yanayin farawa da aka daidaita kuma za a iya maimaitawa** — kafin kowane trial, operators suna daidaita scene da image overlay.
- **Adadin trials masu yawa** — real-world rollouts 50 ga kowane task, policy da condition; 200 ga kowane task a simulation. Jimilla: kusan **blind real-world rollouts 1,800 da simulation rollouts sama da 47,000**.
- **Kididdiga da ta dace** — Bayesian estimates na yiwuwar nasara da pairwise hypothesis tests tare da multiple-comparison corrections, maimakon duban error bars da ido kawai. Sun kuma sake bincika kusan kashi ɗaya cikin huɗu na trials da mutane suka yi scoring domin auna kuskuren scoring.

[video pending]

Kwatantawa gefe da gefe na models suna shirya teburin karin kumallo: a hagu single-task baseline, a dama LBM. Bidiyoyin suna tafiya a 1× speed. Wannan task guda ne da aka tantance, ba hujjar general-purpose autonomy ba.

Wannan tsarin aunawa shi ne muhimmin abu. Gaba ɗayan takardar hujja ce cewa idan ba a yi irin wannan kulawa ba, yana da wahala a bambance **ingantawa ta gaske** da sa'a.

Abin da suka gano

Manyan models da aka yi finetuning sun fi single-task models da aka horar daga farko — a matsakaici. Idan aka haɗa tasks, LBM da aka fara pretraining sannan aka yi finetuning ga task ya fi policy da aka horar daga farko a wannan task, a simulation da real world, kuma bambancin ya kai muhimmancin kididdiga. Idan aka duba tasks ɗaya-ɗaya, finetuned LBM ya kasance statistically as-good-or-better a kusan duk cases: 3/3 real-world tasks da 15/16 simulation tasks.

Babban nasara mafi bayyana ita ce amfani da bayanai da kyau. Finetuned LBM ya kai performance irin na from-scratch model da kusan **sau 3–5 ƙarancin task-specific data**. A wani real-world task — shirya breakfast table — LBM da aka yi finetuning da **15% kawai** na demonstrations ya fi from-scratch policy da aka horar da **100%** na demonstrations.

Pretraining ya fi taimakawa idan yanayi ya kauce daga training. Da aka canza test environment da gangan

daga training conditions — wato **distribution shift** — fa'idar finetuned LBM ta ƙaru. A wani set na simulation, ya fi from-scratch da muhimmancin kididdiga a 3 cikin 16 tasks a normal conditions, amma 10 cikin 16 a distribution shift. Saboda real deployments kusan kullum suna ɗan kaucewa training conditions, wannan robustness na iya zama sakamakon da ya fi muhimmanci a aikace.

Ƙarin pretraining data ya taimaka a hankali. Performance ya ci gaba da hawa yayin da aka ƙara bayanan pretraining, ba tare da wani sudden jump ko "emergent leap" ba a scales da aka gwada. Sakamakon mai amfani ne, mai sauƙin hango direction ɗinsa, kuma ba mai ban mamaki ba.

Amma labarin generalist ba tare da finetuning ba bai samu goyon baya ba. Pretrained LBM da aka yi amfani da shi **zero-shot**, ba tare da task-specific finetuning ba, bai ci single-task policies a kai a kai ba. Network guda zai iya yin tasks da yawa, amma ra'ayin "ka gaya masa kawai abin da kake so" bai fito a nan ba. Marubutan suna ganin ƙaramin language encoder ɗinsu na iya kasancewa wani ɓangare na dalili.

Kuma fa'idodin sun yi ƙanƙanta har ana iya rasa su — ko a ɗauki hayaniya a matsayin sakamako. Yawancin effects sun bayyana ne saboda sample sizes sun fi na al'ada girma kuma tests ɗin sun yi hankali. Marubutan sun ce kai tsaye cewa, idan aka duba girman effects da hayaniyar gwaji, **akwai haɗari mai muhimmanci cewa wasu takardun robotics suna auna statistical noise.** Sun kuma gano cewa wani abu mai sauƙi — yadda ake **normalise data** — ya shafi sakamako fiye da architectural changes, kuma wani normalisation **bug** a pretraining bai bayyana ba sai bayan evaluation ta ƙare.

Abin da wannan yake iya nufi

Karatun da zai iya tsayawa a kan hujja shi ne: large-scale pretraining a kan robot data masu bambance-bambance ingredient ne mai amfani. Yana rage adadin task-specific data da ake buƙata ga sabon aiki kuma yana sa policies su fi jure lokacin da duniya ba ta yi daidai da training ba. Wannan yana goyon bayan direction da fannin yake zuba jari a kai. Amma fa'idodin **matsakaici ne kuma suna da sharadi**: yawanci suna bayyana bayan finetuning, kuma sun fi fitowa idan an haɗa tasks ko lokacin stress. Ba su nuna general robot da za a sauke ya yi komai ba.

Ma'anar da ta fi shiru kuma watakila ta fi muhimmanci ita ce methodological. Takardar kamar ma'aunin aunawa ce: tana nuna yawan shaidar da ake buƙata kafin a yi ikirari mai aminci game da robot policy, kuma tana nuna cewa wani ɓangare na armashin da aka wallafa na iya zama yana kan samples masu ƙanƙanta. Fannin yana buƙatar irin wannan gyara fiye da wani sabon model da ya hau leaderboard.

Abin da wannan binciken bai tabbatar ba

- **Ba general-purpose robot ba ne.** An nuna fa'idar architecture takamaimai — diffusion policies da ake yi wa finetuning ga kowane task — a controlled settings daga teleoperated demonstrations. Ba robot autonomous da zai yi duk sabon aiki da aka gaya masa ba.
- Bai tabbatar da **zero-shot** use ba. Ba tare da finetuning ba, babban model bai fi single-task baselines a kai a kai ba.
- Ba shaida ce ta **"emergent leap"** ba. Scaling ya inganta performance a hankali; babu discontinuity da zai goyi

bayan labarin “sai kwatsam ya zama capable.”

- Lambobin suna da alaƙa da lab. Absolute success rates an daidaita su kusa da 50% da gangan domin comparisons su fi sensitive; ba ma’aunin real-world reliability ba ne. Kuma architecture ɗaya ce daga lab ɗaya.
- Bai warware **me ya sa** policy ɗaya take nasara ko gazawa ba. Wasu specific tasks inda LBM ya yi muni an ruwaito su amma ba a yi musu cikakken bayani ba.
- Bai ce komai game da **safety, autonomy ko deployment** a wajen evaluation rig ba.

Yaya ƙarfin shaidar yake?

Ga manyan comparative claims — finetuned LBMs sun fi from-scratch baselines idan aka haɗa tasks, suna buƙatar sau da yawa ƙarancin data, kuma sun fi robust a distribution shift — shaidar tana da ƙarfi kuma an sarrafa gwajin da kyau: blind, randomized, large-sample, da statistical testing, tare da QA na scoring. A wannan yanayin, methodology ta isa a ɗauki manyan conclusions kusa da yadda aka ruwaito su.

Iyakokin da suka rage su ne waɗanda marubutan suka bayyana da kansu. Error bars ɗinsu suna auna randomness na **evaluation**, amma ba randomness na **training** ba — idan ka horar da model iri ɗaya sau biyu, za ka iya samun policies masu bambanci sosai, kuma wannan variation ba ya cikin statistics. Real-world tasks suna da trials 50 kowanne, wanda ya isa ga medium effects amma na iya rasa small ones. Language conditioning ya yi amfani da modest encoder, don haka sakamakon “just tell the robot what to do” na iya bambanta da manyan systems. Kuma akwai disclosure na normalisation bug da aka gano bayan evaluation. Babu ɗayan waɗannan da ya rushe babban sakamakon, amma su ne irin matsalolin da takardar ke cewa fannin bai kamata ya ɓoye ba.

Wani bayanin source: wannan explainer ya dogara da preprint na marubutan. Ba a samu journal-published version domin a kwatanta ko an yi canje-canje daga preprint ba.

Me ya sa wannan yake da muhimmanci

“Robot foundation models” kalma ce da ta dace da overclaiming, kuma wannan binciken yana da saukin a karanta shi ta hanyoyi biyu marasa kyau: “ya yi aiki!” ko “duk hype ne.” Karatun da ya fi amfani yana tsakani: pretraining a kan diverse data yana ba da **fa’idodi na gaske, masu aunawa amma matsakaici** — musamman ƙarancin data ga sabon task da ƙarin robustness — kuma fa’idar tana ƙaruwa a hankali da scale.

Dalilin da ya fi zurfi shi ne takardar ta juya tsauraran matakan aunawa zuwa fannin kanta. Ta nuna cewa effects na gaske sun yi ƙanƙanta har za su ɓace a sloppy evaluation, kuma boring choice kamar data normalisation na iya zama mafi muhimmanci fiye da clever new architecture. Wannan hujja ce cewa ci gaban robot learning yana buƙatar ingantaccen measurement kafin a yarda da shi. Takardar da take kashe wani ɓangare na armashinta wajen raba sakamako daga fata tana yin abin da ya fi topping a leaderboard daraja.

Takaitaccen bayani

Masu bincike a Toyota Research Institute sun horar da **large behavior models** — diffusion-based robot policies da aka yi pretraining da kusan awa 1,700 na manipulation data masu bambance-bambance — sannan suka gwada su da from-scratch single-task policies ta wata hanya mai tsauri: blind, randomized, large-sample, kusan **real-world rollouts 1,800 da simulation trials sama da 47,000**, tare da kididdiga na gaske. Bayan finetuning ga kowane task, manyan models sun fi kyau idan aka haɗa tasks, sun buƙaci kusan **sau 3–5 ƙarancin task-specific data**, kuma sun fi **robust lokacin da conditions suka canza**, yayin da performance ta ƙaru a hankali da ƙarin pretraining data. Amma **ba tare da finetuning ba** ba su fi single-task models a kai a kai ba. Wasu effects sun yi ƙanƙanta har large sample size ne kawai ya bayyana su, kuma data-normalisation choice mai sauƙi ya fi wasu architectural differences tasiri. Sakamakon goyon baya ne mai ma'auni ga direction na robot foundation models — **ba** general-purpose robot ba, ba zero-shot generalist ba, kuma ba "emergent leap" ba — tare da gargadi cewa wani ɓangare na robotics na iya zama yana auna noise.

Bincike ba tare da ƙarin gishiri ba

Abin da takardar ta nuna: Da blind, randomized, statistically powered evaluation mai kusan real-world rollouts 1,800 da simulation rollouts sama da 47,000, multitask-pretrained sannan finetuned diffusion policies (LBMs) sun fi from-scratch single-task policies idan aka haɗa sakamako, sun kai equivalent performance da kusan sau 3–5 ƙarancin task-specific data, kuma sun fi robust ƙarƙashin distribution shift. Performance ta ƙaru a hankali da pretraining data.

Abin da yake yiwuwa amma ba a tabbatar ba: Waɗannan fa'idodi za su koma manyan vision-language-action models; da smooth scaling zai ci gaba fiye da range na data da aka gwada.

Abin da bai nuna ba: General-purpose ko zero-shot robot; wani "emergent" capability jump; cikakken bayani ga specific task failures; real-world reliability a absolute terms; ko wani abu game da safety da autonomous deployment.

Manyan iyakoki: Statistics suna ɗaukar evaluation randomness amma ba training-run randomness ba; real-world trials 50 ga kowane task na iya rasa small effects; architecture ɗaya da lab ɗaya; modest language encoder; data-normalisation bug da aka gano bayan evaluation; nazarin ya dogara da preprint, ba journal version da aka kwatanta ba.

Yawan amincewar da ya dace ga mai karatu na gama gari: Babba cewa multitask pretraining tare da finetuning yana ba da real, moderate benefits — musamman wajen data efficiency da robustness — kuma an auna su da kulawa da ba kasafai ake gani ba. Babba kuma cewa wannan **ba** general-purpose ko zero-shot robot ba ne kuma **ba** emergent leap ba ne. Matsakaici kan yadda gains ɗin za su ci gaba da scale zuwa manyan models. Kuma gargadin marubutan ya cancanci a ɗauke shi da muhimmanci: effects a robotics na iya zama ƙanana har under-powered studies su ruwaito noise a matsayin progress.

Majiyoyi

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/