



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

Sé àwọn robot “foundation models” nílá sìşé dáadáa jù? Ìdáhùn tó sọra: diẹ ni béèni, ọpọ study kò sì lè sọ iyàtò

Toyota Research Institute kọ “large behavior models” — robot policies tí a pretrain lórí ~1,700 wákàtí diverse manipulation data — wọn sì fi wọn wé from-scratch single-task policies pèlú blind, randomized, large-sample evaluation (~1,800 real-world, 47,000+ simulation) àti statistics gidi. Lẹyìn per-task fine-tuning, big models ẹe dáadáa ju ní average, nílò tó 3–5× task-specific data tó kéré sí i, wọn sì robust sí i nígbà conditions yí padà; performance pò lówólówó bí pretraining data ẹe pò. Sùgbón láìsí finetuning wọn kò consistently ẹgun single-task models, ọpọ effect kéré tó bẹẹ tí sample nílá nìkan fi hàn wọn, data-normalisation choice mundane kan sì ní ipa tó pò ju architecture lọ. Èyí jẹ support tó measured fún robot-foundation-model direction — kì í ẹe general-purpose robot, kì í ẹe zero-shot generalist, kì í ẹe “emergent leap” — pèlú ikilọ pé ọpọ robotics research lè n wọn noise.

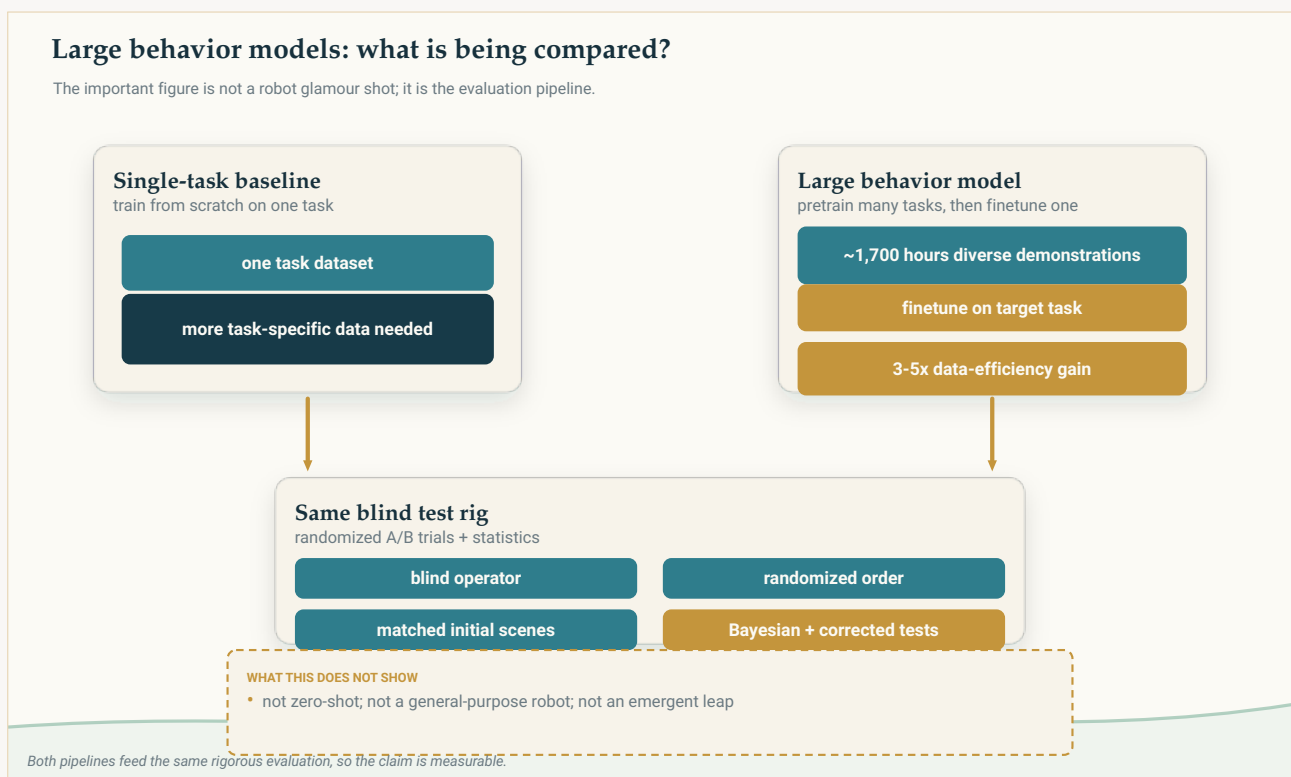
Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

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

ìwé ìwádíí robot kan tí koko gidi rẹ jẹ òtító

Robotics wà láàrín ìfẹ tuntun sí “foundation àwòṣe”. Èrò náà, tí a gba láti èdè àti image AI, ní fa ènìyàn mọra: dípò kí a kọ robot kan fún iṣẹ kan lẹ́ẹkan, kọ **“large ìhùwàsí àwòṣe” (LBM)** nílá kan lórí àkójọpọ demonstration tó pọ, tó sì yàtò sí ara wọn, kí a sì ní ètò tó lè ṣe ọpọ̀ nńkan tí ó sì lè fara mó iṣẹ tuntun kíákíá. Ìtara náà — àti owó tí wọn ní fí sínú rẹ — pọ̀ gan-an. àkọlé rọrùn láti kọ: *ọpọ̀lo robot tó lè ṣe gbogbo nńkan ti dé*.

ìwé ìwádíí yíí láti Toyota Research Institute ṣe pàtàkì gan-an nítorí pé ó kò láti kọ àkọlé yẹn. Contribution gidi rẹ kì í ṣe robot tó ní ṣe spectacle tó pọ̀ sí i. Ó jẹ àyèwò líle sí ìbèèrè tó dà bí ẹni pé ó rọrùn — *ṣe big-model ọ̀nà náà ṣiṣe dáadára jù lọ ní tòótọ́, báwo la sì ṣe lè mó?* — pẹ̀lú ìṣọra statistical tí àwọn olùkòwé fúnra wọn sọ pé kò wọpọ̀ nínú field náà. Abajade jẹ “bèèni, ṣùgbọ̀n” tó wúlò gan-an, àti ìkìlò pé ọpọ̀ robotics research lè jẹ pé wọn ní wọn ariwo.



Approach méjèèjì wọ inú blind, randomized evaluation kan náà. Claim ìwé ìwádíí náà kì í ṣe pé robot foundation àwòṣe jẹ zero-shot generalists, bí kò ṣe pé pretraining lè mú dàtá efficiency àti robustness dára sí i nígbà tí a bá wọn wọn pẹ̀lú ìṣọra.

Original The Clean Paper diagram CC BY 4.0

Ohun tí àwọn olùkòwé ṣe

Wọn kọ LBM ní itumọ̀ tó kedere: **diffusion-based visuomotor policies** — Diffusion Transformer kan tó ka image camera, instruction èdè kukuru kan, àti joint position robot fúnra rẹ, tí ó sì ní dá motor command kékeré jáde ní 10 Hz. Wọn **pretrain wọn lórí tó 1,700 wákàtí** demonstration robot — ju iṣẹ 500 lọ tí wọn kó jọ nínú lab wọn,

pèlú public datasets — lẹyìn nàà wọn **finetune** wọn fún iṣẹ kọọkan. Ifiwéra jálẹ̀ iwé iwádíí jẹ̀ pèlú **single-task policy tí a kọ láti odo** lórí dátà iṣẹ kan nàà.

Sùgbón ọkàn iwé iwádíí nàà ni *evaluation*, tí àwọn olùkòwé ka sí abajade pàtàkì jù lọ. Láti má ẹ tan ara wọn jẹ, wọn lo:

- **Blind, randomized A/B testing** nínú ayé gidi — ẹnì tó n ṣiṣẹ robot nàà kò mọ policy wo ni a n ídánwò, order nàà sì jẹ randomized.
- **Controlled, repeatable initial conditions** — ǻǻǻjú ídánwò kọọkan, operators mú scene nàà bá image overlay mu.
- **ídánwò púpọ̀** — 50 ayé gidi rollout fún iṣẹ-ṣiṣe kọọkan, policy kọọkan àti condition kọọkan; 200 fún iṣẹ-ṣiṣe kọọkan nínú simulation. Lapapọ: tó **1,800 blind ayé gidi rollouts àti ju 47,000 simulation rollouts** lọ.
- **Statistics tó tọ** — Bayesian iṣíró àfojúsùn fún probability of success, àti pairwise hypothesis ídánwò pèlú multiple-comparison corrections, dípò kí wọn kan wo àṣiṣe bars tó overlap. Wọn tún ẹ quality-assurance lórí ìdámẹ́rín àwọn ídánwò tí ènìyàn score láti wọn scoring àṣiṣe.

[video pending]

Àfiwé side-by-side tí àwòṣe méjì tí wọn n Ẹ̀tò tabili breakfast: ní ọ̀sì, single-task baseline; ní ọ̀tún, LBM. Fídíò méjèèjì n ṣiṣẹ ní 1x speed. Èyí jẹ iṣẹ-ṣiṣe kan tí a ẹ evaluate, kí í ẹ ẹrì general-purpose autonomy.

Èrọ evaluation yíí gan-an ni kókó. Gbogbo iwé iwádíí nàà n sọ pé láìsì rẹ, o ọ̀rọ láti mọ ìmúgbòdòrò gidi kúrò nínú orííre.

Ohun tí wọn rí

Big àwòṣe tí a finetune dára ju single-task àwòṣe tí a kọ láti odo lọ — ní average. Nígbà tí wọn kó iṣẹ-ṣiṣe jọ, LBM tí a pretrain, tí a sì finetune fún iṣẹ-ṣiṣe kan, ẹ dáadára ju policy tí a kọ láti odo lórí iṣẹ-ṣiṣe nàà lọ, nínú simulation àti ayé gidi, separation nàà sì jẹ statistically significant. Fún iṣẹ-ṣiṣe kọọkan, finetuned LBM jẹ statistically as-good-or-better ju from-scratch lọ ní fere gbogbo ọ̀ràn (3/3 ayé gidi iṣẹ-ṣiṣe, 15/16 simulation iṣẹ-ṣiṣe).

Èrè tó tóbi jù, tó sì mọ̀ jù, ni dátà efficiency. Finetuned LBM dé performance tó dọgba pèlú from-scratch àwòṣe nípa lílo tó **3–5× dátà task-specific tó kéré sí i**. Nínú ayé gidi iṣẹ-ṣiṣe kan — ṣiṣe tabili breakfast — LBM tí wọn finetune lórí **15%** demonstration nìkan ẹ́gun from-scratch policy tí a kọ lórí **100%** demonstration.

Pretraining ẹ iranlọwọ̀ jù nígbà tí condition bá yí padà. Nígbà tí wọn mọ̀mọ̀ yí ídánwò environment kúrò ní training conditions (“distribution shift”), advantage finetuned LBM *pọ̀ sí i*. Nínú simulation set kan, ó ẹ́gun from-scratch statistically lórí 3 nínú 16 iṣẹ-ṣiṣe ní normal condition, sùgbón 10 nínú 16 ní distribution shift. Nítorí deployment gidi máa n yà sí training conditions, robustness yíí lè jẹ abajade tó wúlò jù lọ ní iṣe.

Pretraining dátà tó pọ̀ sí i ẹ iranlọwọ̀, lówólówọ̀. Performance gòkè diédìe bí wọn ẹ fi pretraining dátà kún un — kò sí sudden jump tàbí “emergent” leap ní scale tí wọn ídánwò. Ó wúlò, ó ẹ́ẹ̀ sọ̀tẹ̀lẹ̀, kò sì ní drama.

Sùgbón itàn generalist láìsì finetuning kò dúró. Pretrained LBM tí a lo ní *zero-shot* — láìsì task-specific

finetuning — kò consistently sẹgun single-task policies. Network kan lè ẹ ọpọ iṣẹ-ṣiṣe, sùgbọ́n àlá “kan fún un ní prompt” kò farahàn níbí; àwọn olùkòwé sọ pé apá kan lè jẹ nítorí èdè encoder kékeré wọn kò lágbára tó.

Àwọn gain nàà sì kéré tó bẹ̀ẹ̀ tí ó rọ̀rùn láti pàdánù wọn — tàbí láti ẹ ẹni pé wọn wà. Ọpọ ipa kò hàn tí tí àpẹrẹ iwọn fi tóbi ju bó ẹ wọpọ lẹ àti tí idánwò fi sọra. Àwọn olùkòwé sọ ní kedere pé, níwò iye ipa àti ariwo, ewu tó pọ wà pé ọpọ robotics iwé iwádii n wọn statistical ariwo. Wọn tún rí pé yíyàn tó dà bí ohun kékeré — bí a ẹ normalize dátà — ní ipa tó pọ ju architectural changes lẹ, àti pé normalization bug kan nínú pretraining kò farahàn tí tí lẹyin tí evaluation parí.

Ohun tí ẹyí ẹ́é túmò sí

Ìtumò tó lè dáàbò bo: large-scale pretraining lórí robot dátà tó yàtò jẹ ingredient gidi tó wúlò — ó dín dátà tuntun tí a nílò fún iṣẹ-ṣiṣe koṣkan kù, ó sì mú policy lágbára sí i nígbà tí ayé gidi kò bá training mu. Ẹyí ẹ ẹ àtílẹ̀yìn fún fún itọ́sọ̀nà tí field ní tẹ sí. Sùgbọ́n gain nàà jẹ moderate àti conditional: wọn hàn jù lẹyin finetuning, ní aggregate, àti nígbà stress; kì í ẹ ibí drop-in gbogbogbò robot kan.

Ìtumò mfi tó dakẹ sùgbọ́n tó ẹ pàtākì jù ni methodology. iwé iwádii nàà dà bí measuring stick: ó fi iye ẹrì tí a nílò gan-an hàn láti sọ claim tó ẹ́é gbẹkẹ lé nípa robot policy, ó sì n daba pé ọpọ excitement tí a tẹ jáde dúró lórí ẹrì tó kéré ju. Field nàà nílò correction yíi ju àwòṣe tuntun mfi lẹ.

Ohun tí ẹyí kò fi ẹrì múlẹ

- Kì í ẹ **general-purpose robot**. Àwọn win nàà wà fún architecture kan pato (diffusion policies), tí a finetune fún iṣẹ-ṣiṣe koṣkan, nínú controlled settings, láti teleoperated demonstrations — kì í ẹ autonomous robot tó lè ẹ iṣẹ tuntun kankan nípa command.
- Kò validate **zero-shot** lẹ. Láisi finetuning, big àwòṣe kò consistently sẹgun single-task baselines.
- Kì í ẹ ẹrì fún “**emergent leap**.” Scaling mú performance dára sí i lẹwọlẹwọ; kò sí discontinuity tí ó ẹ ẹ àtílẹ̀yìn fún fún itàn “lẹyin nàà ó lojiji di capable.”
- Àwọn nọmbà jẹ **relative, lab-bound**. Absolute success rates ni wọn mọmọ mú sún mọ ~50% kí ifiwéra lè jẹ sensitive; kì í ẹ measure ti ayé gidi reliability, iṣe nàà sì jẹ architecture kan láti lab kan.
- Kò ṣàlàyé **idi** tí policy kan fi ṣaṣeyọrí tàbí kuna; iṣe-ṣiṣe kan tó jẹ pé big àwòṣe ẹ burú sí i wà, sùgbọ́n wọn kò ṣàlàyé gbogbo wọn.
- Kò sọ ohunkóhun nípa **ààbò, autonomy, tàbí deployment** níta evaluation rig.

Báwo ni ẹrì ẹ lágbára tó?

Fún central comparative claims — pé finetuned LBM sẹgun from-scratch baselines ní aggregate, nílò dátà tó kéré sí i lẹpọ igbà, ó sì robust sí distribution shift — ẹrì nàà lágbára ó sì controlled ju bó ẹ wọpọ lẹ: blind, randomized, large-sample, statistically tested, pẹlú QA lórí scoring. Ẹyí jẹ ọkan lára àwọn rare ọràn tí methodology dára tó bẹ̀ẹ̀

tí a lè gba àkọlé conclusion rẹ ní pẹkipekí.

Àwọn caveat tó jẹ òtítọ ní àwọn olùkòwé fúnra wọn sọ. Error bars wọn gba randomness ti *evaluation*, sùgbón **kò gba randomness ti training** — tí o bá train àwòṣe kan lèṣe, o lè gba policy tó yàtò ní meaningful way, variation yẹn kò sì wà nínú statistics. ayé gidi iṣe-ṣiṣe ní 50 ìdánwò kọọkan, tó lè rí medium ipa sùgbón tó lè padanu small ipa. Language-conditioning lo encoder tó modest, nítorí náà claim “kan sọ fún robot ohun tí ó fẹ ẹ” lè yà nínú ètò tó tóbi sí i. Wọn tún sọ normalization bug tí wọn rí léyìn evaluation ní kedere. Kò sí èyíkẹyí tó pa pàtàkì àwàrí run, sùgbón wọn jẹ gangan irú nìkan tí ìwé ìwádíí náà sọ pé field sáà maa ní kojá sílẹ.

Àkọsílẹ̀ orisun kan ní èmí kan náà: explainer yíi dá lórí preprint àwọn olùkòwé. A kò lè gba journal-published èdà, nítorí náà a kò ṣàyèwò bóyá ayipada wà láàárín preprint àti published text.

Kí nìdí tí ó fi ẹ pàtàkì?

“Robot foundation àwòṣe” jẹ phrase tó rọrùn láti overclaim, ìwé ìwádíí bá yíi sì rọrùn láti ka lónà méjì tó burú — gégé bí “ó ṣiṣẹ!” tàbí “*overhyped ni gbogbo rẹ*.” Ìtumò tó pé wúlò ju méjèjì lọ: pretraining lórí diverse dátà ní fúnni ní ànfààní gidi, measurable, sùgbón moderate — pàápàá dátà tó kéré fún iṣe-ṣiṣe tuntun àti robustness tó pọ — performance sì ní dara sí i ní ọna tó ṣeé sọtẹlẹ bí scale ẹ ní pọ.

Ìdí tó jinlẹ̀ jù ní pé ìwé ìwádíí náà lo rigor rẹ sí field tirẹ. Nípa fífi hàn pé ipa gidi kéré tó bẹ̀ẹ̀ tí sloppy evaluation lè mú un parẹ, àti pé yíyàn mundane bí dátà normalization lè ní ipa tó pọ̀ ju architecture tuntun ọlọgbọn lọ, ó ní jìyan pé ọpọ̀ robot-learning progress nílò measurement tó lágbára sí i kí a tó gbà á gbọ. ìwé ìwádíí tó ní lo credibility tirẹ̀ láti sọ ààlà láàárín abajade àti ifẹ̀ jẹ ohun tó sọwọn, tó sì ní iye ju jíjẹ number ọkan lórí leaderboard lọ.

Àkótán kedere

Àwọn olùwádíí ní Toyota Research Institute kọ “large ìhùwàsí àwòṣe” — diffusion-based robot policies tí a pretrain lórí tó ~1,700 wákàtí diverse manipulation dátà — wọn sì fi wọn wé from-scratch single-task policies pèlú protocol tó rigor gan-an: blind, randomized, large-sample ($\approx 1,800$ ayé gidi àti 47,000+ simulation ìdánwò), pèlú statistics gidi. Léyìn per-task finetuning, big àwòṣe ẹ daádáa ju ní aggregate, wọn nílò tó **3–5× dátà task-specific tó kéré sí i**, wọn sì **robust sí i nígbà tí conditions yí padà**, performance sì ní gòkè lówọlówọ bí pretraining dátà ẹ pọ. Sùgbón **làisí finetuning**, wọn kò consistently sẹgun single-task àwòṣe; ọpọ̀ ipa kéré tó bẹ̀ẹ̀ tí àpẹrẹ̀ nílá nìkan fi hàn wọn, data-normalization choice mundane kan sì ní ipa tó pọ̀ ju architecture lọ. Èyí jẹ ẹ àtílẹ̀yìn fún tó measured fún ìtọ́sọ̀nà robot-foundation-model — **kì í ẹ general-purpose robot, kì í ẹ zero-shot generalist, kì í ẹ “emergent leap”** — pèlú ìkílò tó tó pé ọpọ̀ robotics research lè jẹ statistical ariwo.

Àyèwò láisí àṣejù

Ohun tí ìwé ìwádíí fi hàn: Pèlú blind, statistically-powered evaluation tó rigor ($\approx 1,800$ ayé gidi + 47,000+ simulation rollouts), multitask-pretrained-then-finetuned diffusion policies (LBMs) sẹgun from-scratch single-

task policies ní aggregate, dé performance tó dógba pèlú $\sim 3\text{--}5\times$ task-specific dátà tó kéré sí i, wọn sì robust sí distribution shift; performance ní pò lówólówó pèlú pretraining dátà.

Ohun tó ẹẹ gba sùgbón tí a kò fi ẹrí múlẹ: Pé ànfààní wònyí yóò gbe lọ sí vision-language-action àwòṣe tó tóbi gan-an (èdè encoder wọn kéré); pé smooth scaling náà yóò tẹ síwájú ju ààlà dátà tí wọn ìdánwò lọ.

Ohun tí kò fi hàn: General-purpose tàbí zero-shot robot; “emergent” capability jump; ìdí fún gbogbo task-level ìkùnà; absolute ayé gidi reliability (success rates ni wọn mú sún mọ 50% fún sensitivity); ohunkóhun nípa ààbò tàbí autonomous deployment.

Main limitations: Statistics gba evaluation randomness sùgbón kì í gba training-run randomness; 50 ayé gidi ìdánwò fún iṣẹ-ṣiṣe lè padanu small ipa; architecture kan àti lab kan; èdè encoder modest; data-normalization bug kan ni wọn rí léyìn evaluation; itúpalẹ dá lórí preprint, published èdà kò tii jẹ checked.

Confidence wo ni gbogbogbò reader yẹ kí ó ní? Gíga pé multitask pretraining + finetuning ní fúnni ní ànfààní gidi, moderate — pàápàá dátà efficiency àti robustness — àti pé wọn wọn wọn pèlú ìṣọra tó sòwọn. Gíga pé èyí **kì í ẹ general-purpose tàbí zero-shot robot, kì í sì ẹ emergent leap.** Medium lórí bí gain náà ẹ maa scale sí àwòṣe tó tóbi. Ó sì yẹ ká gba ìkìlò àwọn olùkòwé ní sérieux: ipa field náà kéré tó bẹẹ tí under-powered iwádíi lè maa ròyìn ariwo. Ìpò tò yẹ: measured optimism nípa ọ̀nà náà, pèlú skepticism tó dára sí robot-AI èsì tí kò ní irú statistical backing yí.

Àwọn orísun

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/>