The Clean Paper

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

Ṣé àwọn robot "foundation models" ńlá ṣiṣẹ́ dáadáa jù? Ìdáhùn tó ṣọ́ra: díẹ̀ ni béẹ̀ni, ọ̀pọ̀ study kò sì lè sọ ìyàtọ̀

Toyota Research Institute kọ "large behavior models" — robot policies tí a pretrain lórí ~1,700 wákàtí diverse manipulation data — wọ́n sì fi wọn wé from-scratch single-task policies pẹ̀lú blind, randomized, large-sample evaluation (~1,800 real-world, 47,000+ simulation) àti statistics gidi. Lẹ́yìn per-task fine-tuning, big models ṣe dáadáa ju ní average, nílò tó 3–5× task-specific data tó kéré sí i, wọ́n sì robust sí i nígbà conditions yí padà; performance pọ̀ lọ́wọ́lọ́wọ́ bí pretraining data ṣe pọ̀. Ṣùgbọ́n láìsí finetuning wọn kò consistently ṣẹ́gun single-task models, ọ̀pọ̀ effect kéré tó bẹ́ẹ̀ tí sample ńlá nìkan fi hàn wọn, data-normalisation choice mundane kan sì ní ipa tó pọ̀ ju architecture lọ. Èyí jẹ́ support tó measured fún robot-foundation-model direction — kì í ṣe general-purpose robot, kì í ṣe zero-shot generalist, kì í ṣe "emergent leap" — pẹ̀lú ìkìlọ̀ pé ọ̀pọ̀ robotics research lè ń wọn noise.

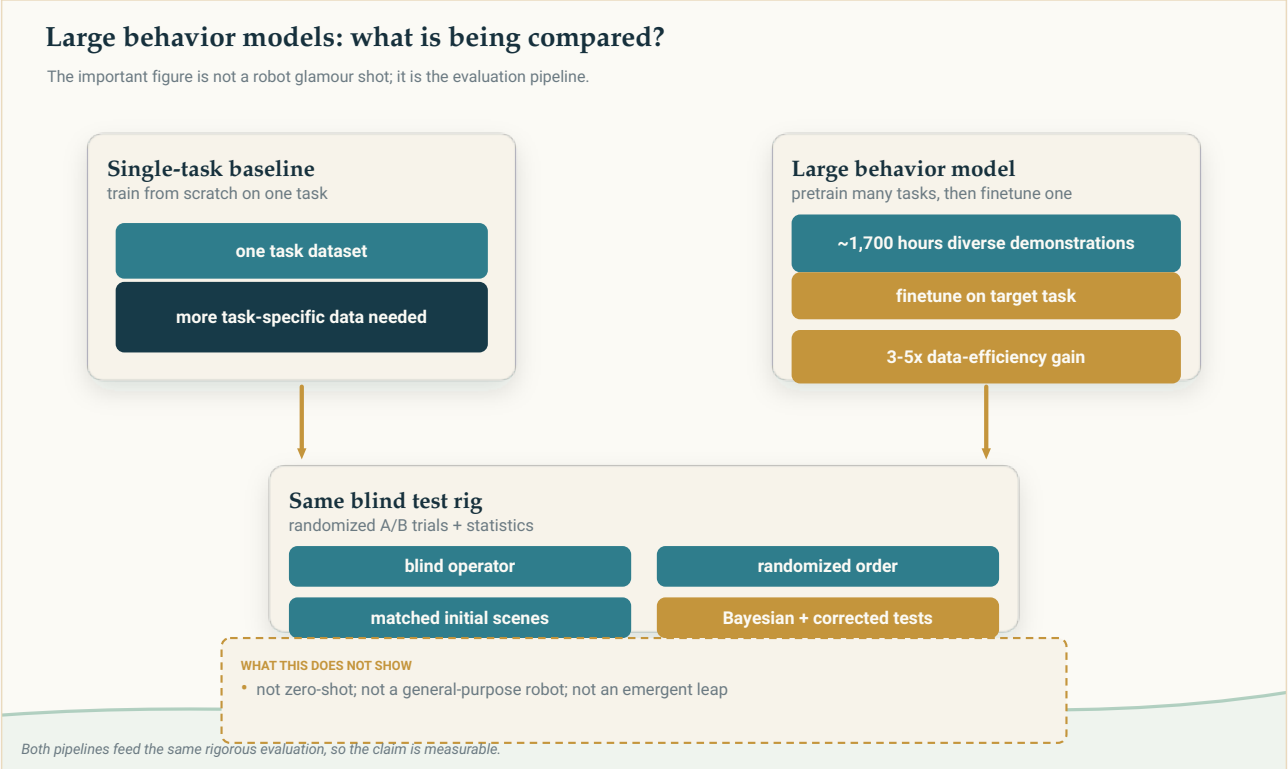
Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

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 ìwádìí robot kan tí koko gidi rẹ jẹ́ òtítọ́

Robotics wà láàrín ìfẹ́ tuntun sí “foundation àwòṣe”. Èrò náà, tí a gba láti èdè àti image AI, ń fa ènìyàn mọ́ra: dípò kí a kọ robot kan fún iṣẹ́ kan lẹ́ẹ̀kan, kọ **“large ìhùwàsí àwòṣe” (LBM)** nlá kan lórí àkójọpọ̀ demonstration tó pọ̀, tó sì yàtọ̀ sí ara wọn, kí a sì ní ètò tó lè ṣe ọ̀pọ̀ nǹkan tí ó sì lè fara mọ́ iṣẹ́ tuntun kíákíá. Ìtara náà — àti owó tí wọ́n ń fi sínú rẹ̀ — pọ̀ gan-an. àkọlé rọrùn láti kọ: *ọpọlọ robot tó lè ṣe gbogbo nǹkan ti dé.*

ìwé ìwádìí yìí láti Toyota Research Institute ṣe pàtàkì gan-an nítorí pé ó kọ̀ láti kọ àkọlé yẹn. Contribution gidi rẹ̀ kì í ṣe robot tó ń ṣe spectacle tó pọ̀ sí i. Ó jẹ́ àyẹ̀wò líle sí ìbéèrè tó dà bí ẹni pé ó rọrùn — *ṣé big-model ọ̀nà náà ṣiṣẹ́ dáadáa jù lọ ní tòótọ́, báwo la sì ṣe lè mọ̀?* — pẹ̀lú ìṣọ́ra statistical tí àwọn olùkọ̀wé fúnra wọn sọ pé kò wọ́pọ̀ nínú field náà. Abajade jẹ́ “béẹ̀ni, ṣùgbọ́n” tó wúlò gan-an, àti ìkìlọ̀ pé ọ̀pọ̀ robotics research lè jẹ́ pé wọ́n ń wọn ariwo.



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

Ohun tí àwọn olùkọ̀wé ṣe

Wọ́n 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 rẹ̀, tí ó sì ń dá motor command kékeré jáde ní 10 Hz. Wọ́n **pretrain wọn lórí tó 1,700 wákàtí** demonstration robot — ju iṣẹ́ 500 lọ tí wọ́n kó jọ nínú lab wọn,

pẹ̀lú public datasets — léyìn náà wọ́n **finetune** wọn fún iṣẹ́ kọọkan. ìfiwéra jálẹ̀ ìwé ìwádìí jẹ́ pẹ̀lú **single-task policy tí a kọ́ láti odo** lórí dátà iṣẹ́ kan náà.

Ṣùgbọ́n ọkàn ìwé ìwádìí náà ni *evaluation*, tí àwọn olùkọ̀wé ka sí abajade pàtàkì jù lọ. Láti má ṣe tan ara wọn jẹ, wọ́n lo:

- **Blind, randomized A/B testing** nínú ayé gidi — ẹni tó ń ṣiṣẹ́ robot náà kò mọ policy wo ni a ń ìdánwò, order náà sì jẹ́ randomized.
- **Controlled, repeatable initial conditions** — ṣáájú ìdánwò kọọkan, operators mú scene náà bá image overlay mu.
- **ìdánwò púpọ̀** — 50 ayé gidi rollout fún iṣẹ́-ṣiṣe kọọkan, policy kọọkan àti condition kọọkan; 200 fún iṣẹ́-ṣiṣe kọọkan nínú simulation. Lapapọ: tó **1,800 blind ayé gidi rollouts àti ju 47,000 simulation rollouts lọ**.
- **Statistics tó tọ́** — Bayesian ìṣírò àfojúsùn fún probability of success, àti pairwise hypothesis ìdánwò pẹ̀lú multiple-comparison corrections, dípò kí wọ́n kan wo àṣìṣe bars tó overlap. Wọ́n tún ṣe quality-assurance lórí ìdámẹ́rin àwọn ìdánwò tí ènìyàn score láti wọn scoring àṣìṣe.

[video pending]

Àfiwé side-by-side ti àwòṣe méjì tí wọ́n ń ṣètò tabili breakfast: ní òsì, single-task baseline; ní ọ̀tún, LBM. Fídíò méjèèjì ń ṣiṣẹ́ ní 1x speed. Èyí jẹ́ iṣẹ́-ṣiṣe kan tí a ṣe evaluate, kì í ṣe èrí general-purpose autonomy.

Ẹ̀rọ evaluation yìí gan-an ni kókó. Gbogbo ìwé ìwádìí náà ń sọ pé láìsí rẹ̀, o ṣòro láti mọ ìmúgbòòrò gidi kúrò nínú oríire.

Ohun tí wọ́n rí

Big àwòṣe tí a finetune dára ju single-task àwòṣe tí a kọ́ láti odo lọ — ní average. Nígbà tí wọ́n kó iṣẹ́-ṣiṣe jọ, LBM tí a pretrain, tí a sì finetune fún iṣẹ́-ṣiṣe kan, ṣe dáadáa ju policy tí a kọ́ láti odo lórí iṣẹ́-ṣiṣe náà lọ, nínú simulation àti ayé gidi, separation náà sì jẹ́ statistically significant. Fún iṣẹ́-ṣiṣe kọọkan, finetuned LBM jẹ́ statistically as-good-or-better ju from-scratch lọ ní fere gbogbo ọ̀ràn (3/3 ayé gidi iṣẹ́-ṣiṣe, 15/16 simulation iṣẹ́-ṣiṣe).

Èrè tó tóbi jù, tó sì mọ́ jù, ni dátà efficiency. Finetuned LBM dé performance tó dọ́gba pẹ̀lú from-scratch àwòṣe nípa lílo tó **3–5× dátà task-specific tó kéré sí i**. Nínú ayé gidi iṣẹ́-ṣiṣe kan — ṣíṣe tabili breakfast — LBM tí wọ́n finetune lórí **15%** demonstration nìkan ṣẹ́gun from-scratch policy tí a kọ́ lórí **100%** demonstration.

Pretraining ṣe iranlọwọ́ jù nígbà tí condition bá yí padà. Nígbà tí wọ́n mọ̀ọ́mọ̀ yí ìdánwò environment kúrò ní training conditions (“distribution shift”), advantage finetuned LBM *pọ̀ sí i*. Nínú simulation set kan, ó ṣẹ́gun from-scratch statistically lórí 3 nínú 16 iṣẹ́-ṣiṣe ní normal condition, ṣùgbọ́n 10 nínú 16 ní distribution shift. Nítorí deployment gidi máa ń yà sí training conditions, robustness yìí lè jẹ́ abajade tó wúlò jù lọ ní iṣe.

Pretraining dátà tó pọ̀ sí i ṣe iranlọwọ́, lọ́wọ́lọ́wọ́. Performance gòkè díẹ̀díẹ̀ bí wọ́n ṣe fi pretraining dátà kún un — **kò sí sudden jump tàbí “emergent” leap** ní scale tí wọ́n ìdánwò. Ó wúlò, ó ṣeé sọ tẹ́lẹ̀, kò sì ní drama.

Ṣùgbọ́n ìtàn generalist láìsí finetuning kò dúró. Pretrained LBM tí a lo ní *zero-shot* — láìsí task-specific

finetuning — kò consistently ṣ́égun single-task policies. Network kan lè ṣe òpọ̀ iṣ́é-ṣiṣe, ṣùgbọ́n àlá “kan fún un ní prompt” kò farahàn níbí; àwọn olùkọ̀wé sọ pé apá kan lè jẹ́ nítorí èdè encoder kékeré wọn kò lágbára tó.

Àwọn gain náà sì kéré tó bẹ́ẹ̀ tí ó rọrùn láti pàdánù wọn — tàbí láti ṣe ẹni pé wọ́n wà. Òpọ̀ ipa kò hàn títí tí àpẹẹrẹ ìwọ̀n fi tóbi ju bó ṣe wọ́pọ̀ lọ àti tí ìdánwò fi ṣọ́ra. Àwọn olùkọ̀wé sọ ní kedere pé, níwò iye ipa àti ariwo, **ewu tó pọ̀ wà pé òpọ̀ robotics ìwé ìwádìí ń wọn statistical ariwo.** Wọ́n tún rí pé yíyàn tó dà bí ohun kékeré — bí a ṣe *normalize* dátà — ní ipa tó pọ̀ ju architectural changes lọ, àti pé normalization **bug** kan nínú pretraining kò farahàn títí *lẹ́yìn* tí evaluation parí.

Ohun tí èyí ṣeé túmọ̀ sí

Ìtumọ̀ tó lè dáàbò bo: large-scale pretraining lórí robot dátà tó yàtọ̀ jẹ́ ingredient gidi tó wúlò — ó dín dátà tuntun tí a nílò fún iṣ́é-ṣiṣe kọọkan kù, ó sì mú policy lágbára sí i nígbà tí ayé gidi kò bá training mu. Èyí ṣe ṣe àtìlẹ́yìn fún fún ìtọ́ṣọ́nà tí field ń tẹ̀ sí. Ṣùgbọ́n gain náà jẹ́ *moderate àti conditional*: wọ́n hàn jù lọ lẹ́yìn finetuning, ní aggregate, àti nígbà stress; kì í ṣe ìbí drop-in gbogbogbò robot kan.

Ìtumọ̀ mîì tó dakẹ́ ṣùgbọ́n tó ṣe pàtàkì jù ni methodology. ìwé ìwádìí náà dà bí measuring stick: ó fi iye ẹ̀rí tí a nílò gan-an hàn láti sọ claim tó ṣeé gbẹ́kẹ̀ lé nípa robot policy, ó sì ń daba pé òpọ̀ excitement tí a tẹ̀ jáde dúró lórí ẹ̀rí tó kéré ju. Field náà nílò correction yìí ju àwòṣe tuntun mîì lọ.

Ohun tí èyí kò fi ẹ̀rí múlẹ̀

- Kì í ṣe **general-purpose robot.** Àwọn win náà wà fún architecture kan pato (diffusion policies), tí a finetune fún iṣ́é-ṣiṣe kọọkan, nínú controlled settings, láti teleoperated demonstrations — kì í ṣe autonomous robot tó lè ṣe iṣ́é tuntun kankan nípa command.
- Kò validate **zero-shot** lò. Láìsí finetuning, big àwòṣe kò consistently ṣ́égun single-task baselines.
- Kì í ṣe ẹ̀rí fún **“emergent leap.”** Scaling mú performance dára sí i lọ́wọ́lọ́wọ́; kò sí discontinuity tí ó ṣe ṣe àtìlẹ́yìn fún fún ìtàn “lẹ́yìn náà ó lojiji di capable.”
- Àwọn nọ́mbà jẹ́ **relative, lab-bound.** Absolute success rates ni wọn mọ̀ọ́mọ̀ mú sún mọ́ ~50% kí ìfiwéra lè jẹ́ sensitive; kì í ṣe measure ti ayé gidi reliability, iṣ́é náà sì jẹ́ architecture kan láti lab kan.
- Kò ṣàlàyé **ìdí** tí policy kan fi ṣaṣeyọrí tàbí kuna; iṣ́é-ṣiṣe kan tó jẹ́ pé big àwòṣe ṣe burú sí i wà, ṣùgbọ́n wọn kò ṣàlàyé gbogbo wọn.
- Kò sọ ohunkóhun nípa **ààbò, autonomy, tàbí deployment** níta evaluation rig.

Báwo ni ẹ̀rí ṣe lágbára tó?

Fún central comparative claims — pé finetuned LBM ṣ́égun from-scratch baselines ní aggregate, nílò dátà tó kéré sí i lọ́pọ̀ ìgbà, ó sì robust sí distribution shift — ẹ̀rí náà lágbára *ó sì controlled ju bó ṣe wọ́pọ̀ lọ*: blind, randomized, large-sample, statistically tested, pẹ̀lú QA lórí scoring. Èyí jẹ́ òkan lára àwọn rare òràn tí methodology dára tó bẹ́ẹ̀

tí a lè gba àkọlé conclusion rẹ ní pẹkipẹki.

Àwọn caveat tó jẹ òtítọ ni àwọn olùkọ̀wé fúnra wọn sọ. Error bars wọn gba randomness ti *evaluation*, ṣùgbọ́n **kò gba randomness ti training** — tí o bá train àwòṣe kan lẹ́ẹ̀mejì, o lè gba policy tó yàtọ̀ ní meaningful way, variation yẹn kò sì wà nínú statistics. ayé gidi iṣẹ́-ṣiṣe ní 50 ìdánwò kọọkan, tó lè rí medium ipa ṣùgbọ́n tó lè padanu small ipa. Language-conditioning lo encoder tó modest, nítorí náà claim "kan sọ fún robot ohun tí ó fẹ́ ṣe" lè yà nínú ètò tó tóbi sí i. Wọ́n tún sọ normalization bug tí wọ́n rí lẹ́yìn evaluation ní kedere. Kò sí èyíkéyìí tó pa pàtàkì àwárí run, ṣùgbọ́n wọ́n jẹ́ gangan irú nǹkan tí ìwé ìwádìí náà sọ pé field sábà máa ń kọjá sílẹ̀.

Àkọsílẹ̀ orísun kan ní ẹ̀mí kan náà: explainer yìí dá lórí preprint àwọn olùkọ̀wé. A kò lè gba journal-published ẹ̀dà, nítorí náà a kò ṣàyẹ̀wò bóyá ayipada wà láàárín preprint àti published text.

Kí nìdí tí ó fi ṣe pàtàkì?

"Robot foundation àwòṣe" jẹ́ phrase tó rọrùn láti overclaim, ìwé ìwádìí báyìí sì rọrùn láti ka lọ́nà méjì tó burú — gẹ́gẹ́ bí *"ó ṣiṣẹ́!"* tàbí *"overhyped ni gbogbo rẹ̀."* Ìtumọ̀ tó pé wúlò ju méjèèjì lọ: pretraining lórí diverse dátà ń fúnni ní àǹfààní gidi, measurable, ṣùgbọ́n moderate — pàápàá dátà tó kéré fún iṣẹ́-ṣiṣe tuntun àti robustness tó pọ̀ — performance sì ń dara sí i ní ọna tó ṣeé sọtẹ́lẹ̀ bí scale ṣe ń pọ̀.

Ìdí tó jinlẹ̀ jù ni pé ìwé ìwádìí náà lo rigor rẹ̀ sí field tirẹ̀. Nípa fífi hàn pé ipa gidi kéré tó bẹ́ẹ̀ tí sloppy evaluation lè mú un parẹ́, àti pé yíyàn mundane bí dátà normalization lè ní ipa tó pọ̀ ju architecture tuntun ọlọ́gbọ́n lọ, ó ń jiyan pé ọ̀pọ̀ robot-learning progress nílò measurement tó lágbára sí i kí a tó gbà á gbọ́. ìwé ìwádìí tó ń lo credibility tirẹ̀ láti ṣọ ààlà láàárín abajade àti ìfẹ́ jẹ́ ohun tó ṣọ̀wọ́n, tó sì ní iye ju jíjẹ́ number ọ̀kan lórí leaderboard lọ.

Àkótán kedere

Àwọn olùwádìí ní Toyota Research Institute kọ "large ìhùwàsí àwòṣe" — diffusion-based robot policies tí a pretrain lórí tó ~1,700 wákàtí diverse manipulation dátà — wọ́n sì fi wọn wé from-scratch single-task policies pẹ̀lú protocol tó rigor gan-an: blind, randomized, large-sample (≈1,800 ayé gidi àti 47,000+ simulation ìdánwò), pẹ̀lú statistics gidi. Lẹ́yìn per-task finetuning, big àwòṣe ṣe dáadáa ju ní aggregate, wọ́n nílò tó **3–5× dátà task-specific tó kéré sí i**, wọ́n sì **robust sí i nígbà tí conditions yí padà**, performance sì ń gòkè lọ́wọ́lọ́wọ́ bí pretraining dátà ṣe pọ̀. Ṣùgbọ́n **láìsí finetuning**, wọn kò consistently ṣégun single-task àwòṣe; ọ̀pọ̀ ipa kéré tó bẹ́ẹ̀ tí àpẹẹrẹ ńlá nìkan fi hàn wọn, data-normalization choice mundane kan sì ní ipa tó pọ̀ ju architecture lọ. Èyí jẹ́ ṣe àtìlẹ́yìn fún tó measured fún ìtọ́ṣọ́nà robot-foundation-model — **kì í ṣe general-purpose robot, kì í ṣe zero-shot generalist, kì í ṣe "emergent leap"** — pẹ̀lú ìkìlọ̀ tó tọ́ pé ọ̀pọ̀ robotics research lè jẹ́ statistical ariwo.

Àyẹ̀wò láìsí àṣejù

Ohun tí ìwé ìwádìí fi hàn: Pẹ̀lú blind, statistically-powered evaluation tó rigor (≈1,800 ayé gidi + 47,000+ simulation rollouts), multitask-pretrained-then-finetuned diffusion policies (LBMs) ṣégun from-scratch single-

task policies ní aggregate, dé performance tó dọ́gba pẹ̀lú ~3–5× task-specific dátà tó kéré sí i, wọ́n sì robust sí distribution shift; performance ń pọ̀ lọ́wọ́lọ́wọ́ pẹ̀lú pretraining dátà.

Ohun tó ṣeé gbà ṣùgbọ́n tí a kò fi ẹ̀rí múlẹ̀: Pé àìnfàànì wọ̀nyí yóò gbe lọ sí vision-language-action àwòṣe tó tóbi gan-an (èdè encoder wọn kéré); pé smooth scaling náà yóò tẹ̀ síwájú ju ààlà dátà tí wọ́n 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 wọ́n mú sún mọ́ 50% fún sensitivity); ohunkóhun nípa ààbò tàbí autonomous deployment.

Main limitations: Statistics gba evaluation randomness ṣùgbọ́n kì í gba training-run randomness; 50 ayé gidi ìdánwò fún iṣẹ́-ṣiṣe lè padanu small ipa; architecture kan àti lab kan; èdè encoder modest; data-normalization bug kan ni wọ́n rí lẹ́yìn evaluation; ìtúpalẹ̀ dá lórí preprint, published ẹ̀dà kò tíì jẹ́ checked.

Confidence wo ni gbogbogbò reader yẹ kí ó ní? Gíga pé multitask pretraining + finetuning ń fúnni ní àìnfàànì gidi, moderate — pàápàá dátà efficiency àti robustness — àti pé wọ́n wọn wọ́n pẹ̀lú ìṣọ́ra tó ṣọ̀wọ́n. Gíga pé èyí **kì í ṣe general-purpose tàbí zero-shot robot, kì í sì ṣe emergent leap**. Medium lórí bí gain náà ṣe máa scale sí àwòṣe tó tóbi. Ó sì yẹ ká gba ìkìlọ̀ àwọn olùkọ̀wé ní sérieux: ipa field náà kéré tó bẹ́ẹ̀ tí under-powered ìwádìí lè máa ròyìn ariwo. Ìpò tó yẹ: measured optimism nípa ọ̀nà náà, pẹ̀lú skepticism tó dára sí robot-AI èsì tí kò ní irú statistical backing yìí.

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/