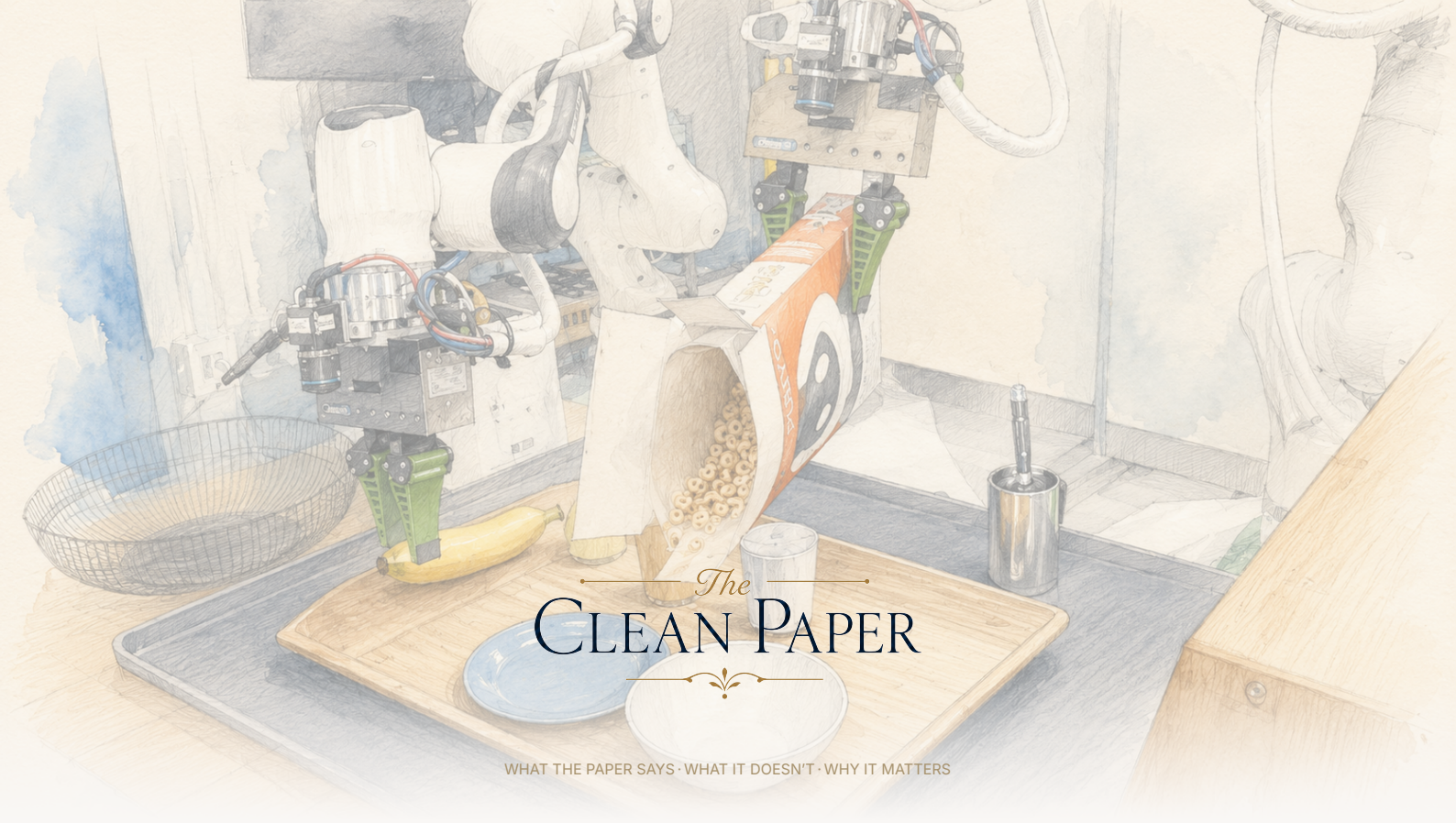


The Clean Paper

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

Дали големите роботски „foundation models“ навистина работат подобро? Внимателниот одговор е: умерено да — а многу студии не можат ни да го видат тоа

Toyota Research Institute обучи „large behavior models“ — роботски политики pre-trained на ~1,700 часа разновидни манипулациски податоци — и ги тестираше против single-task политики обучени од нула со невообичаена строгост: слепи, рандомизирани, големи примероци (~1,800 реални, 47,000+ симулациски) со вистинска статистика. По finetuning по задача, големите модели во просек беа подобри, бараа приближно 3–5× помалку task-specific податоци и беа поробусни кога условите се менуваа; перформансите растеа мазно со повеќе pretraining data. Но без finetuning не ги победуваа конзистентно single-task моделите, неколку ефекти беа толку мали што бараа големи примероци, а баналeн избор на нормализација значеше повеќе од архитектурата. Тоа е измерена поддршка за robot-foundation-model насоката — не општ робот, не zero-shot генералист, не „emergent скок“ — и предупредување дека многу роботика можеби мери шум.

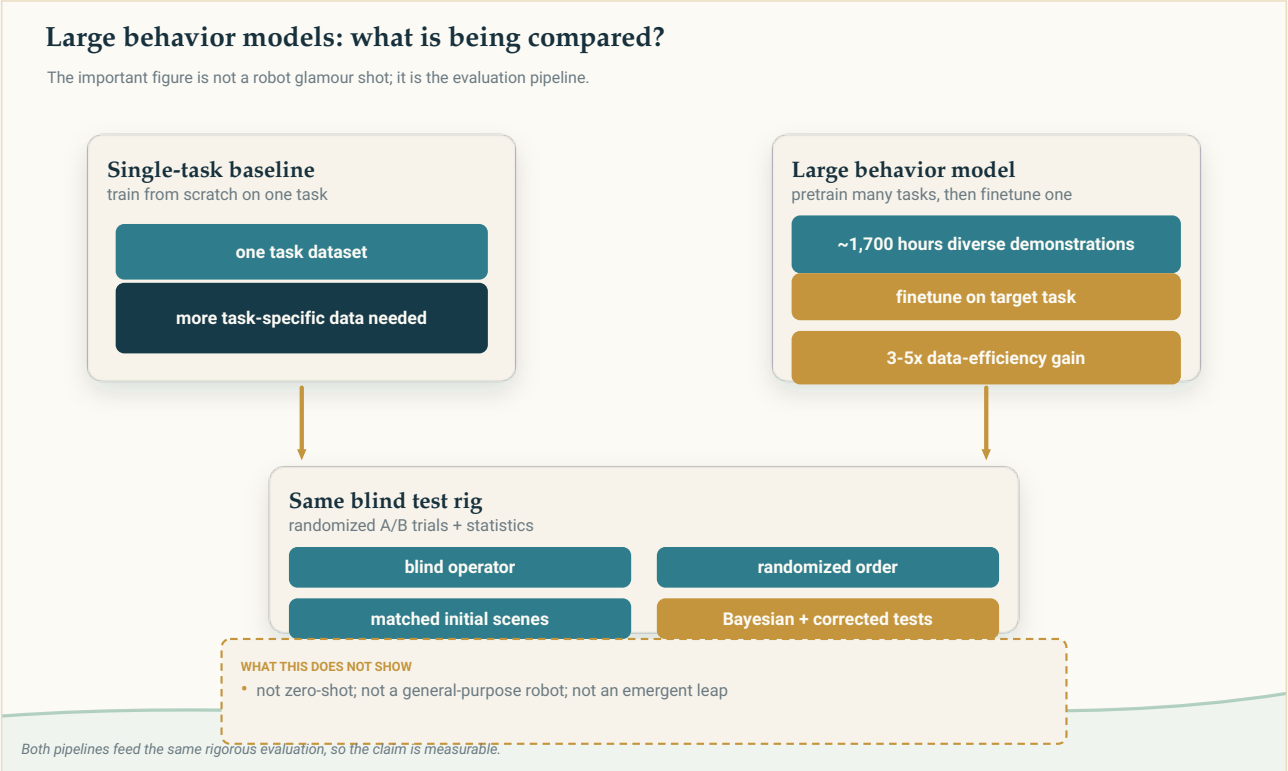
Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

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

Труд за роботи чија вистинска тема е чесното мерење

Роботиката е среде трка по „foundation models“. Идејата, позајмена од јазичната и сликовната AI, е примамлива: наместо робот да се обучува за една задача по една, обучете еден голем **„large behavior model“ (LBM)** на огромен и разновиден куп демонстрации и добијте систем со широки способности кој брзо се приспособува. Ентузијазмот — и инвестициите — се огромни. Насловот речиси сам се пишува: *пристигнаа општите мозоци за роботи.*

Овој труд од Toyota Research Institute е интересен токму затоа што одбива да го напише тој наслов. Неговиот вистински придонес не е поспектакуларен робот. Тоа е строг поглед на навидум едноставно прашање — *дали пристапот со голем модел навистина работи подобро и како воопшто би знаеле?* — одговорено со ниво на статистичка грижа што, според самите автори, е невообичаено за полето. Резултатот е навистина корисно „да, но“, плус предупредување дека голем дел од роботиката можеби мери шум.



Двата пристапа влегуваат во иста слепа, рандомизирана евалуација. Тврдењето на трудот не е дека robot foundation models се zero-shot генералисти, туку дека pretraining може да ја подобри ефикасноста на податоците и робусноста кога се мери внимателно.

Што направиле авторите

Тие изградиле LBM во конкретна смисла: **diffusion-based visuomotor политики** (Diffusion Transformer што чита слики од камера, кратка јазична инструкција и положби на зглобовите на роботот, а потоа испорачува кратки серии моторни команди на 10 Hz). Моделите биле **pretrained на приближно 1,700 часа** роботски демонстрации — повеќе од 500 различни задачи собрани внатрешно, плус јавни datasets — и потоа **finetuned** за поединечни задачи. Споредбата низ трудот е со **single-task политика обучена од нула** само на податоците за таа задача.

Но срцето на трудот е *евалуацијата*, која авторите ја третираат како главен резултат. За да не се измамат самите себе, користеле:

- **Слеп, рандомизиран A/B тест** во реалниот свет — човекот што го пуштал роботот не знаел која политика се тестира, а редоследот бил рандомизиран.
- **Контролирани, повторливи почетни услови** — операторите ја усогласувале сцената со слика-overlay пред секој обид.
- **Голем број обиди** — 50 реални rollouts по задача, политика и услов; 200 по задача во симулација. Вкупно: околу **1,800 слепи реални rollouts и повеќе од 47,000 симулациски rollouts**.
- **Соодветна статистика** — Bayesian проценки на веројатноста за успех и pairwise тестови на хипотези со корекции за повеќекратни споредби, наместо визуелно споредување преклопени error bars. Дури спровеле QA проверка на четвртина од човечки оценетите обиди за да ја измерат грешката во оценувањето.

[video pending]

Споредба една до друга на модели што поставуваат маса за појадок: лево single-task baseline, десно LBM. Двата видеа се пуштаат со 1× брзина. Ова е една евалуирана задача, не доказ за општа автономија.

Оваа мерна машинерија е поентата. Целиот труд е аргумент дека без неа не можете да разликувате вистинско подобрување од среќа.

Што откриле

Finetuned големите модели ги победуваат single-task моделите обучени од нула — во просек. Кога резултатите се агрегираат низ задачи, LBM што бил pretrained па finetuned за конкретната задача сигурно ја надминувал политиката обучена од нула на истите task-specific податоци, и во симулација и во реалниот свет, со статистички значајна разлика. На ниво на поединечни задачи, finetuned LBM бил статистички ист или подобар речиси секогаш (3/3 реални задачи, 15/16 симулациски).

Најголемата и најчистата добивка е ефикасноста на податоците. Finetuned LBM

достигнувал перформанси еквивалентни на from-scratch политика со приближно **3–5× помалку податоци за конкретната задача**. Во една реална задача — поставување маса за појадок — LBM finetuned со само **15%** од демонстрациите ја победил from-scratch политиката обучена со **100%**.

Pretraining помага најмногу кога условите се менуваат. Кога тест-средината намерно била оддалечена од условите на обука („distribution shift“), предноста на finetuned LBM *растела*. Во еден симулациски сет, статистички победувал from-scratch на 3 од 16 задачи во нормални услови, но на 10 од 16 под distribution shift. Бидејќи реалните deployments неизбежно отстапуваат од обуката, оваа робусност веројатно е најпрактично важниот резултат.

Повеќе pretraining податоци помагале постепено. Перформансите растеле стабилно со додавање pretraining data — **без ненадеен скок или „emergent“ момент** на испитаните размери. Корисно, предвидливо, недраматично.

Но приказната за генералист без finetuning не се потврдила. Pretrained LBM користен *zero-shot* — без task-specific finetuning — не ги победувал конзистентно single-task политиките. Една мрежа можела да прави многу задачи, но сонот „само кажи му што да направи“ не се реализирал тука; авторите дел од проблемот го припишуваат на кршливоста на нивниот мал language encoder.

И добивките биле доволно мали лесно да се пропуштат — или случајно да се „пронајдат“. Многу ефекти станале видливи само со невообичаено големите примероци и внимателните тестови. Авторите отворено велат дека, со оглед на големината на ефектите и шумот, **постои значаен ризик многу трудови во роботиката да мерат статистички шум.** Нашле и дека една банална одлука — како се *нормализираат* податоците — влијаела повеќе од архитектонски промени, а bug во нормализацијата за pretraining бил откриен дури *по* завршувањето на евалуациите.

Што најверојатно значи ова

Одбранливото читање: pretraining во голем размер на разновидни роботски податоци е реално и корисно — бара помалку податоци за секоја нова задача и ги прави политиките поотпорни кога светот не изгледа точно како обуката. Тоа навистина ја поддржува насоката во која полето вложува. Но добивките се *умерени и условни* — главно се појавуваат по finetuning и најјасни се во агрегат и под стрес — не се пристигнување на готов општ робот.

Потивкото и поважно значење е методолошко. Трудот е, практично, мерна летва: покажува колку доказ навистина е потребен за доверливо тврдење за роботска политика и сугерира дека голем дел од објавениот ентузијазам се потпира на премалку. На полето му треба ваква корекција повеќе од уште еден модел.

Што ова не докажува

- **Не е општ робот.** Победите се демонстрирани за конкретна архитектура (diffusion policies), finetuned по задача, во контролирани услови, од teleoperated демонстрации — не автономен робот што извршува произволни нови задачи по команда.
- **Не** ја валидира **zero-shot** употребата. Без finetuning, големиот модел не ги победува конзистентно single-task baseline-ите.
- **Не** е доказ за **„emergent скок“**. Скалирањето ги подобрува резултатите мазно; нема дисконтинуитет што би поддржал приказна „и одеднаш стана способен“.
- Бројките се **релативни и лабораториски**. Апсолутните success rates намерно биле наместени околу ~50% за споредбите да бидат чувствителни; тие не се мерка на реална сигурност, а работата е една архитектура од една лабораторија.
- **Не** објаснува *зошто* конкретна политика успева или не, а неколку задачи каде големиот модел бил *полош* се пријавени, но не и објаснети.
- Не кажува **ништо за безбедност, автономија или deployment** надвор од евалуацискиот rig.

Колку се убедливи доказите?

За централните споредбени тврдења — finetuned LBM ги победуваат from-scratch baseline-ите во агрегат, бараат неколку пати помалку податоци и се поробусни под distribution shift — доказите се силни *и* невообичаено добро контролирани: слепи, рандомизирани, со големи примероци, статистички тестирани и со QA на оценувањето. Ова е редок случај каде методологијата е доволно добра за главните заклучоци да се земат речиси по номинална вредност.

Чесните ограничувања се оние што самите автори ги посочуваат. Error bars ја опфаќаат случајноста на *евалуацијата*, но **не** и случајноста на *обуката* — ако истиот модел се обучи двапати, може да се добијат значајно различни политики, а таа варијација не е во статистиката. Реалните задачи имале 50 обиди секоја, доволно за средни ефекти, но можат да пропуштат мали. Јазичното условување користело скромен encoder, па тврдењата за „само кажи му што да прави“ може да изгледаат поинаку кај поголеми системи. И тука е искреното откривање на bug во нормализацијата по евалуацијата. Ниту едно од овие не ги руши главните резултати, но се токму видот проблеми што трудот вели дека полето често ги турка настрана.

Една забелешка за изворот, во ист дух: ова објаснување се темели на preprint-от на авторите. Не успеавме да ја добиеме journal-published верзијата, па не проверивме дали има промени меѓу preprint и финалниот текст.

Зошто е важно

„Robot foundation models“ е израз создаден за претерување, а ваков труд лесно се чита во двете погрешни насоки — како триумфално *„работи!“* или како отфрлачко *„пренадувано е“*. Точната верзија е покорисна: pretraining на разновидни податоци дава реални, мерливи, но умерени придобивки — главно помалку податоци по задача и поголема робусност — а патот се подобрува предвидливо со скалата.

Подлабоката причина зошто е важен е што трудот ја применува строгоста врз сопственото поле. Покажува дека вистинските ефекти се доволно мали да исчезнат под лоша евалуација и дека здодевна одлука како нормализација на податоци може да надвладее паметна нова архитектура. Така гради аргумент дека голем дел од напредокот во robot learning бара постабилно мерење пред да му се верува. Труд што ја троши сопствената кредибилност на разликување резултат од желба прави нешто поретко — и повредно — од освојување leaderboard.

Кратко резиме

Истражувачите од Toyota Research Institute обучиле „large behavior models“ — diffusion-based роботски политики pretrained на ~1,700 часа разновидни манипулациски податоци — и ги тестирале против single-task политики обучени од нула со невообичаено строг протокол: слепи, рандомизирани, големи примероци (≈1,800 реални и 47,000+ симулациски обиди) со вистинска статистика. По finetuning за секоја задача, големите модели сигурно биле подобри во агрегат, барале приближно **3–5× помалку task-specific податоци** и биле **поробусни кога условите се менувале**, а перформансите растеле мазно со повеќе pretraining data. Но **без finetuning** не ги победувале конзистентно single-task моделите, неколку ефекти биле доволно мали да станат видливи само со големи примероци, а баналeн избор на нормализација влијаел повеќе од архитектурата. Ова е цврста, измерена поддршка за насоката на robot foundation models — **не** општ робот, не zero-shot генералист и не „emergent скок“ — плус директно предупредување дека голем дел од роботиката можеби мери шум.

Проверка без претерување

Што покажува трудот: Со строг, слеп и статистички добро напојуван протокол (≈1,800 реални + 47,000+ симулациски rollouts), multitask-pretrained па finetuned diffusion policies (LBM) ги надминуваат from-scratch single-task политиките во агрегат, достигнуваат еквивалентни перформанси со ~3–5× помалку податоци по задача и се поробусни под distribution shift; перформансите растат мазно со pretraining data.

Што е веројатно, но не е докажано: Овие придобивки се пренесуваат на многу

поголеми vision-language-action модели (нивниот language encoder бил мал); мазното скалирање продолжува и над испитаниот опсег на податоци.

Што не покажува: Општ или zero-shot робот (без finetuning → нема конзистентна предност); каков било „emergent“ скок; објаснување за конкретни неуспеси по задача; реална сигурност во апсолутни термини (success rates биле наместени околу 50% за чувствителност); ништо за безбедност или автономен deployment.

Главни ограничувања: Статистиката ја опфаќа случајноста на евалуацијата, но не и варијацијата меѓу training runs; 50 реални обиди по задача може да пропуштат мали ефекти; една архитектура и една лабораторија; скромен language encoder; bug во нормализацијата откриен по евалуациите; анализата се темели на preprint (финалната објавена верзија не е проверена).

Колкава доверба треба да има општ читател? Висока дека multitask pretraining плус finetuning дава реални, умерени придобивки — особено за ефикасност на податоци и робусност — и дека тие се измерени невообичаено внимателно. Висока дека ова **не е** општ или zero-shot робот и **не е** emergent скок. Умерена за тоа колку далеку добивките ќе скалираат. И вреди сериозно да се земе предупредувањето на авторите дека ефектите во полето се доволно мали недоволно напојувани студии да можат да пријавуваат шум. Соодветен став: измерен оптимизам за пристапот и здрава скепса кон robot-AI резултати без ваква статистичка поддршка.

Извори

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/