I grandi “foundation model” per robot funzionano davvero meglio? Una risposta rigorosa — sì, modestamente, e molti studi non riescono nemmeno a dirlo

Toyota Research Institute ha addestrato “large behavior models” — policy robotiche preaddestrate su circa 1.700 ore di dati di manipolazione diversi — e le ha confrontate con policy single-task addestrate da zero usando un protocollo insolitamente rigoroso: test ciechi, randomizzati, con molti campioni (circa 1.800 prove reali e oltre 47.000 in simulazione) e statistiche vere. Dopo il finetuning per singolo compito, i modelli grandi hanno fatto meglio in media, hanno richiesto circa 3–5× meno dati specifici del compito e sono stati più robusti quando le condizioni cambiavano; le prestazioni sono cresciute gradualmente con più dati di preaddestramento. Ma usati senza finetuning non hanno battuto in modo consistente i modelli single-task, diversi effetti erano abbastanza piccoli da richiedere grandi campioni per essere visti, e una scelta banale di normalizzazione dei dati contava più dell’architettura. È un supporto misurato alla direzione dei robot foundation model — non un robot general-purpose, non un generalista zero-shot, non un “salto emergente” — più un avvertimento netto: molta robotica potrebbe stare misurando rumore.

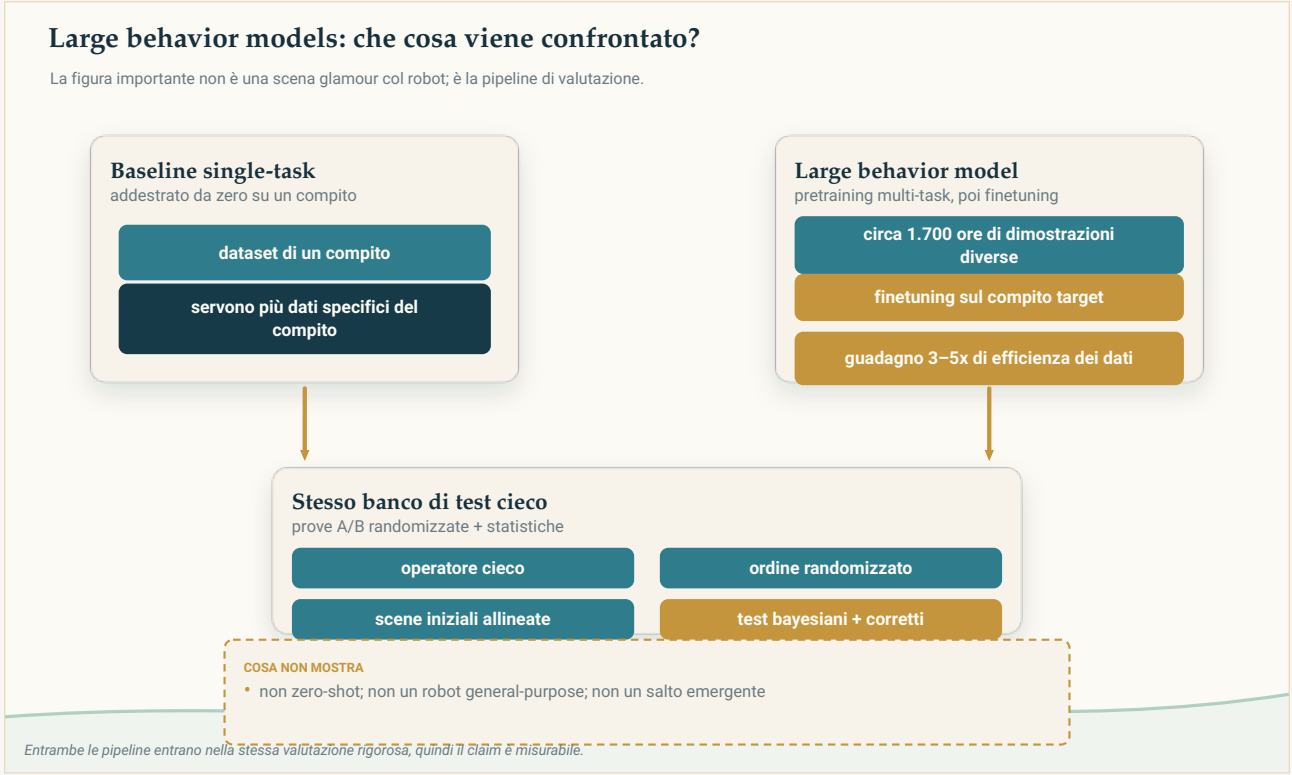
Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

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

Un lavoro sui robot il cui vero tema è l'onestà

La robotica è nel pieno della corsa ai **foundation model**. L'idea, presa in prestito dall'AI per linguaggio e immagini, è seducente: invece di addestrare un robot per un compito alla volta, si addestra un grande **modello comportamentale (*large behavior model*, LBM)** su una pila enorme e varia di dimostrazioni, ottenendo un sistema ampiamente capace e rapido da adattare. L'entusiasmo — e l'investimento — è enorme. Il titolo si scrive da solo: *sono arrivati i cervelli robotici generalisti.*

Questo lavoro, del Toyota Research Institute, è interessante proprio perché rifiuta quel titolo. Il suo vero contributo non è un robot più appariscente. È uno sguardo duro a una domanda ingannevolmente semplice — *l'approccio del modello grande funziona davvero meglio, e come potremmo saperlo?* — con una cura statistica che, per ammissione degli autori, è insolita nel campo. Il risultato è un "sì, ma" realmente utile, e un avvertimento: molta robotica potrebbe stare misurando rumore.



Entrambi gli approcci entrano nella stessa valutazione cieca e randomizzata. L'affermazione del lavoro non è che i modelli robotici di base siano generalisti in modalità zero-shot, ma che il preaddestramento può migliorare l'efficienza dei dati e la robustezza quando viene misurato con cura.

Original The Clean Paper diagram CC BY 4.0

Cosa hanno fatto gli autori

Hanno costruito LBM in un senso specifico e concreto: **politiche di controllo visuomotorie basate su diffusione** (*policy*) (un Diffusion Transformer che legge immagini da telecamera, una breve

istruzione linguistica e le posizioni articolari del robot, poi produce brevi sequenze di comandi motori a 10 Hz). Queste politiche di controllo sono state **preaddestrate su circa 1.700 ore** di dimostrazioni robotiche — oltre 500 compiti distinti raccolti internamente, più set di dati pubblici — e poi sottoposte a **fine-tuning** sui singoli compiti. Il confronto, in tutto il lavoro, è con una **politica di controllo a singolo compito addestrata da zero** sui dati di quel singolo compito.

Il cuore del lavoro, però, è la *valutazione*, che gli autori trattano come il risultato principale. Per evitare di ingannarsi hanno usato:

- **Test A/B ciechi e randomizzati** nel mondo reale — la persona che eseguiva il test non sapeva quale modello fosse in prova, e l'ordine era randomizzato.
- **Condizioni iniziali controllate e ripetibili** — gli operatori allineavano la scena a un'immagine sovrapposta prima di ogni prova.
- **Grandi conteggi di prove** — 50 esecuzioni reali per compito, per modello e per condizione; 200 per compito in simulazione. In totale: circa **1.800 esecuzioni reali in cieco e oltre 47.000 esecuzioni in simulazione**.
- **Statistiche appropriate** — stime bayesiane della probabilità di successo e test di ipotesi a coppie con correzioni per confronti multipli, invece di guardare a occhio barre d'errore sovrapposte. Hanno perfino fatto un controllo qualità su un quarto delle prove valutate da umani per misurare l'errore nell'assegnazione dei punteggi.

[video pending]

Confronto affiancato tra modelli che apparecchiano una tavola per la colazione: a sinistra il modello di riferimento a singolo compito, a destra l'LBM. Entrambi i video sono riprodotti a velocità 1x. È un singolo compito valutato, non una prova di autonomia generale.

Questo apparato è il punto. L'intero lavoro è un argomento a favore dell'idea che, senza tutto questo, non si possa distinguere un miglioramento reale dalla fortuna.

Cosa hanno trovato

Dopo il fine-tuning, i modelli grandi superano in media quelli a singolo compito addestrati da zero. Aggregando i compiti, un LBM preaddestrato e poi sottoposto a fine-tuning su un compito ha superato in modo affidabile un modello addestrato da zero sugli stessi dati del compito, sia in simulazione sia nel mondo reale, con una separazione statisticamente significativa. Sui singoli compiti, l'LBM dopo il fine-tuning è stato statisticamente pari o migliore quasi sempre (3/3 compiti reali, 15/16 in simulazione).

Il vantaggio più grande e chiaro è l'efficienza dei dati. Un LBM dopo il fine-tuning ha raggiunto prestazioni equivalenti al modello da zero usando circa **3–5× meno dati specifici del compito**. In un compito reale (apparecchiare una tavola per la colazione), un LBM sottoposto a fine-tuning con appena il **15%** delle dimostrazioni ha battuto un modello addestrato da zero sul **100%**.

Il preaddestramento aiuta di più quando le condizioni cambiano. Quando l'ambiente di test

veniva perturbato deliberatamente rispetto alle condizioni di addestramento (**cambiamento di distribuzione**, *distribution shift*), il vantaggio dell'LBM dopo il fine-tuning *cresceva*. In un set di simulazione, in condizioni normali batteva statisticamente il modello da zero in 3 compiti su 16, ma in presenza di un cambiamento di distribuzione in 10 su 16. Poiché l'impiego nel mondo reale si discosta sempre, almeno in parte, dalle condizioni di addestramento, questa robustezza è probabilmente il risultato più importante dal punto di vista pratico.

Più dati di preaddestramento aiutavano, in modo graduale. Le prestazioni salivano stabilmente man mano che aggiungevano dati di preaddestramento — senza un salto improvviso o "emergente" alle scale testate. Utile, prevedibile, poco spettacolare.

Ma la storia del generalista senza fine-tuning non ha retto. Un LBM preaddestrato usato *zero-shot* — senza fine-tuning specifico del compito — non ha battuto in modo consistente i modelli a singolo compito. Una singola rete poteva fare molti compiti contemporaneamente, ma il sogno del "basta chiederglielo" non è stato confermato qui; gli autori attribuiscono parte del problema alla fragilità del loro piccolo encoder linguistico.

E i guadagni erano abbastanza piccoli da poter essere mancati — o simulati. Molti effetti sono diventati visibili solo con campioni più grandi del solito e test accurati. Gli autori dicono esplicitamente che, date le dimensioni degli effetti e il rumore, esiste un **rischio significativo che molti lavori di robotica stiano misurando rumore statistico**. Hanno anche trovato che una scelta banale — come i dati vengono *normalizzati* — influenzava i risultati più di cambiamenti architetturali, e che un **bug** di normalizzazione nel preaddestramento è emerso solo *dopo* la fine delle valutazioni.

Cosa probabilmente significa

La lettura difendibile: il preaddestramento su larga scala con dati robotici diversi è un ingrediente reale e utile — fa servire meno dati per ogni nuovo compito e rende le politiche di controllo più robuste quando il mondo non coincide con le condizioni di addestramento. Questo supporta davvero la direzione su cui il campo sta scommettendo. Ma i guadagni sono *moderati e condizionali* (si vedono soprattutto dopo il fine-tuning, e sono più chiari in aggregato e sotto stress), non l'arrivo di un robot generalista pronto all'uso.

Il significato più silenzioso e più importante è metodologico. Il lavoro è, di fatto, un metro di misura: mostra quante prove servano davvero per sostenere un'affermazione credibile su una politica di controllo robotica, e implica che molto entusiasmo pubblicato poggia su troppo poco. È un correttivo di cui il campo ha più bisogno che di un altro modello in cima a una classifica.

Cosa questo non dimostra

- **Non è un robot generalista.** I successi sono mostrati per un'architettura specifica (politiche di controllo basate su diffusione) sottoposta a fine-tuning per compito, in condizioni controllate, a

partire da dimostrazioni teleoperate — non per un robot autonomo che esegue lavori arbitrari a comando.
- **Non** valida l'uso **zero-shot**. Senza fine-tuning, il modello grande non ha battuto in modo consistente i modelli di riferimento a singolo compito.
- **Non** è evidenza di un **"salto emergente"**. L'aumento di scala ha migliorato le prestazioni in modo graduale; non c'è una discontinuità a supporto delle narrazioni "e poi all'improvviso è diventato capace".
- I numeri sono **relativi e di laboratorio**. I tassi di successo assoluti sono stati deliberatamente regolati attorno al 50% per rendere i confronti sensibili; non sono una misura dell'affidabilità nel mondo reale, e il lavoro riguarda una sola architettura in un solo laboratorio.
- **Non** chiarisce **perché** una politica di controllo riesca o fallisca, e diversi compiti specifici in cui il modello grande è andato *peggio* sono riportati ma non spiegati.
- Non dice **nulla su sicurezza, autonomia o impiego reale** fuori dal rig di valutazione.

Quanto sono solide le prove?

Per le affermazioni comparative centrali — gli LBM, dopo il fine-tuning, superano in aggregato i modelli di riferimento addestrati da zero, richiedono diverse volte meno dati e sono più robusti in presenza di un cambiamento di distribuzione — le prove sono solide *e* insolitamente ben controllate: test ciechi, randomizzati, con molti campioni, statisticamente verificati, più un controllo di qualità sull'assegnazione dei punteggi. È un raro caso in cui la metodologia è abbastanza solida da prendere le conclusioni principali quasi alla lettera.

Le cautele più importanti sono quelle sollevate dagli stessi autori. Le loro barre d'errore catturano la casualità della *valutazione*, ma **non** la variabilità dell'*addestramento* — addestrare due volte lo stesso modello potrebbe produrre politiche di controllo significativamente diverse, e questa variazione non entra nelle statistiche. I compiti reali avevano 50 prove ciascuno, abbastanza per cogliere effetti medi ma potenzialmente insufficienti per effetti piccoli. Il condizionamento linguistico usava un encoder modesto, quindi le affermazioni sul "basta dire al robot cosa fare" potrebbero cambiare con sistemi più grandi. E c'è la segnalazione esplicita di un bug di normalizzazione trovato post hoc. Nessuno di questi punti affonda i risultati principali, ma sono esattamente il tipo di cose che, secondo il lavoro, il campo spesso mette da parte.

Una nota sulle fonti, nello stesso spirito: questo explainer si basa sul preprint degli autori. Non siamo riusciti a recuperare la versione pubblicata su rivista, quindi non abbiamo controllato eventuali cambiamenti tra preprint e testo pubblicato.

Perché conta

"Robot foundation model" è una formula che si presta facilmente alle esagerazioni, e uno studio come questo è facile da leggere male in entrambe le direzioni — come un trionfale *"funziona!"* o come un

liquidatorio *"è tutto hype"*. La lettura accurata è più utile di entrambe: il preaddestramento su dati diversi produce benefici reali, misurabili ma moderati — soprattutto meno dati per compito e più robustezza — e il percorso migliora prevedibilmente con la scala.

La ragione più profonda per cui conta è che il lavoro rivolge il suo rigore contro il proprio campo. Mostrando che gli effetti genuini sono abbastanza piccoli da sparire in valutazioni deboli, e che una scelta noiosa come la normalizzazione dei dati può pesare più di una nuova architettura intelligente, costruisce il caso che molti progressi nell'apprendimento robotico abbiano bisogno di misure più robuste prima di essere credibile. Un lavoro che spende la propria credibilità per sorvegliare la differenza tra un risultato e un desiderio sta facendo qualcosa di più raro, e più prezioso, che salire in una classifica.

In breve

I ricercatori del Toyota Research Institute hanno addestrato "large behavior models" — politiche di controllo robotiche basate su diffusione, preaddestrate su circa 1.700 ore di dati diversi di manipolazione — e le hanno testate con modelli a singolo compito addestrati da zero usando un protocollo insolitamente rigoroso: cieco, randomizzato, con molti campioni (circa 1.800 prove reali e oltre 47.000 in simulazione), e statistiche vere. Dopo il fine-tuning per compito, i modelli grandi hanno fatto meglio in aggregato, hanno raggiunto prestazioni equivalenti con circa **3–5× meno dati specifici del compito** e sono stati **più robusti quando le condizioni cambiavano**, con prestazioni che miglioravano gradualmente all'aumentare dei dati di preaddestramento. Ma usati **senza fine-tuning** non hanno battuto in modo consistente i modelli a singolo compito, diversi effetti erano così piccoli che solo i grandi campioni li rendevano visibili, e una scelta banale di normalizzazione dei dati contava più dell'architettura. È un supporto solido e misurato alla direzione dei foundation model per la robotica — **non** un robot generalista, non un generalista zero-shot e non un "salto emergente" — insieme a un avvertimento netto: molta robotica potrebbe stare misurando rumore.

Valutazione critica

Cosa mostra il lavoro: Con una valutazione rigorosa, cieca e statisticamente potente (circa 1.800 esecuzioni reali e oltre 47.000 in simulazione), politiche di controllo basate su diffusione, preaddestrate su più compiti e poi sottoposte a fine-tuning (LBM) superano in aggregato modelli a singolo compito addestrati da zero, raggiungono prestazioni equivalenti con ~3–5× meno dati specifici del compito e sono più robuste in presenza di un cambiamento di distribuzione; le prestazioni scalano gradualmente con i dati di preaddestramento.

Cosa è plausibile ma non dimostrato: Che questi benefici si trasferiscano a modelli visione-linguaggio-azione molto più grandi (il loro encoder linguistico era piccolo); che l'aumento graduale di scala continui oltre l'intervallo di dati testato.

Cosa non mostra: Un robot generalista o zero-shot (senza fine-tuning → nessun vantaggio consis-

tente); qualunque salto “emergente” di capacità; spiegazioni per i fallimenti a livello di singolo compito; affidabilità reale in termini assoluti (i tassi di successo erano tarati vicino al 50% per sensibilità); nulla su sicurezza o impiego autonomo.

Limiti principali: Le statistiche catturano la casualità della valutazione ma non quella delle sessioni di addestramento; 50 prove reali per compito possono mancare effetti piccoli; una sola architettura e un solo laboratorio; encoder linguistico modesto; un bug di normalizzazione dei dati è stato trovato dopo le valutazioni; analisi basata sul preprint (versione pubblicata non controllata).

Quanta fiducia dovrebbe avere un lettore generale? Alta che il preaddestramento multicompito più fine-tuning dia benefici reali e moderati — soprattutto efficienza dei dati e robustezza — e che questi siano stati misurati con cura insolita. Alta che questo **non** sia un robot generalista o zero-shot e **non** un salto emergente. Media su quanto questi guadagni si estendano a modelli più grandi. E da prendere sul serio: l’avvertimento degli autori che gli effetti del campo sono abbastanza piccoli da rendere plausibile che studi sottodimensionati riportino rumore. Atteggiamento appropriato: ottimismo misurato sull’approccio, e sano scetticismo verso risultati sull’AI per la robotica che non hanno questo tipo di supporto statistico.

Fonti

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/