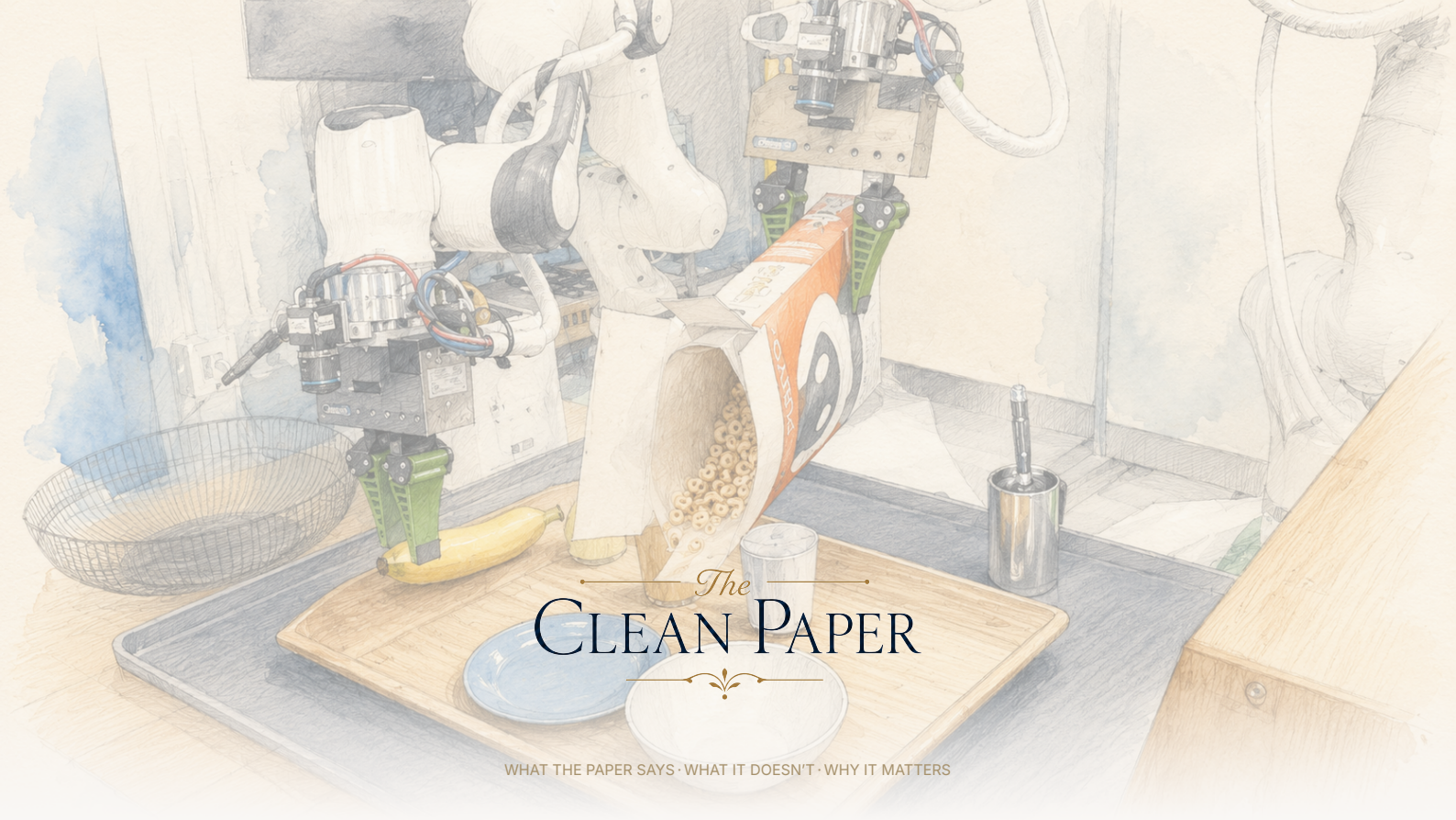


Funktionieren große Roboter-„Foundation-Models“ wirklich besser? Eine sorgfältige Antwort — bescheiden ja, und die meisten Studien können es nicht sagen

Das Toyota Research Institute trainierte „Large Behavior Models“ — Roboter-Policies, vortrainiert auf ~1.700 Stunden vielfältiger Manipulationsdaten — und testete sie gegen von Grund auf trainierte Einzelaufgaben-Policies mit ungewöhnlicher Strenge: blinde, randomisierte Versuche mit großer Stichprobe (~1.800 real, über 47.000 in Simulation) und echter Statistik. Nach aufgabenweisem Finetuning schnitten die Großmodelle im Schnitt besser ab, brauchten rund 3–5× weniger aufgabenspezifische Daten und waren robuster bei verschobenen Bedingungen; die Leistung stieg mit mehr Vortrainingsdaten gleichmäßig. Aber ohne Finetuning schlugen sie Einzelaufgaben-Modelle nicht konsistent, mehrere Effekte waren klein genug, dass erst die großen Stichproben sie sichtbar machten, und eine banale Datennormalisierungs-Entscheidung zählte mehr als die Architektur. Es ist maßvolle Stütze für die Richtung der Roboter-Foundation-Models — kein Universalroboter, kein Zero-Shot-Generalist, kein „emergenter Sprung“ — plus eine pointierte Warnung, dass ein Großteil der Robotik womöglich Rauschen misst.

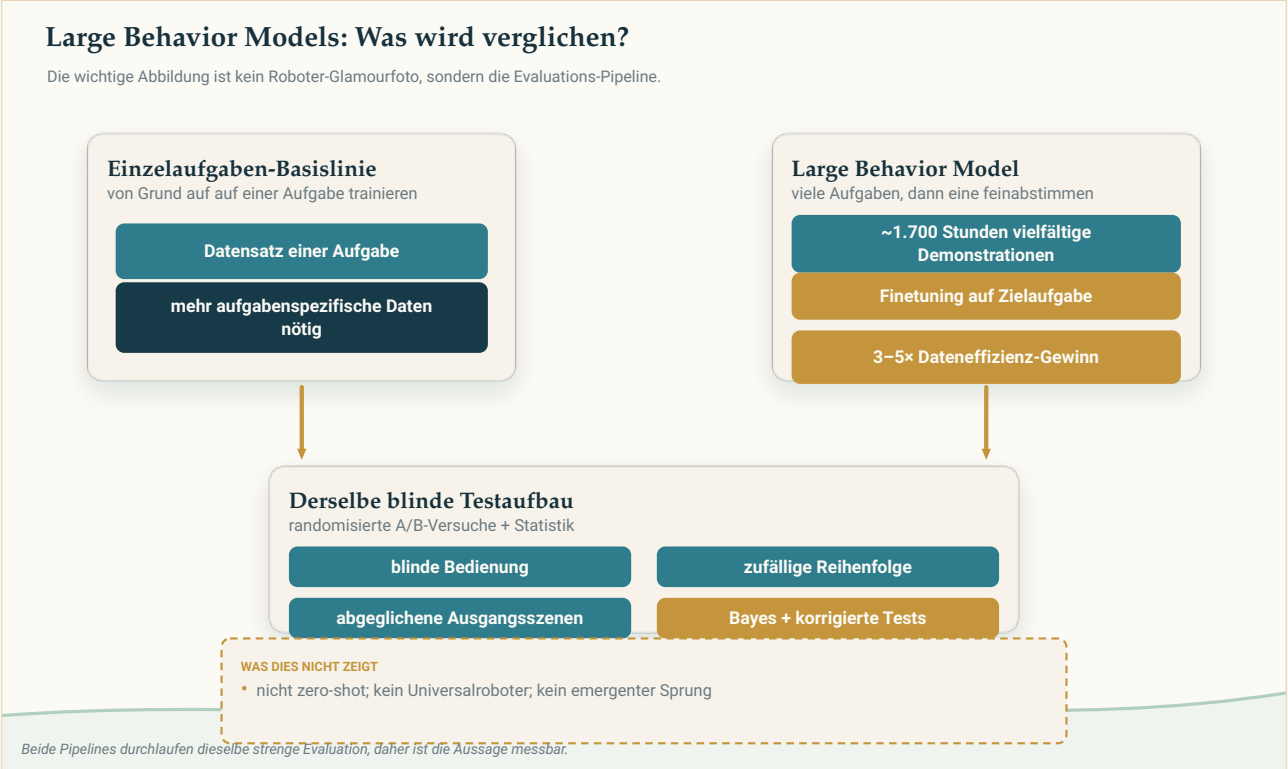
Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

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

Ein Roboter-Paper, dessen eigentliches Thema die Ehrlichkeit ist

Die Robotik steckt mitten in einem „Foundation-Model"-Ansturm. Die Idee, von der Sprach- und Bild-KI entlehnt, ist verführerisch: Statt einen Roboter Aufgabe für Aufgabe zu trainieren, trainiert man ein großes **„Large Behavior Model" (LBM)** auf einem riesigen, vielfältigen Haufen von Demonstrationen und erhält ein System, das breit fähig und schnell anpassungsfähig ist. Die Begeisterung — und die Investitionen — sind enorm. Die Schlagzeile schreibt sich von selbst: *Universelle Roboterhirne sind da.*

Dieses Paper vom Toyota Research Institute ist gerade deshalb interessant, weil es sich weigert, diese Schlagzeile zu schreiben. Sein eigentlicher Beitrag ist kein schnittigerer Roboter. Es ist ein nüchterner Blick auf eine täuschend einfache Frage — *funktioniert der Großmodell-Ansatz tatsächlich besser, und woran würden wir das überhaupt erkennen?* —, beantwortet mit einer statistischen Sorgfalt, die, nach eigener Aussage der Autoren, für das Feld ungewöhnlich ist. Das Ergebnis ist ein wirklich nützliches „ja, aber" und eine Warnung, dass ein Großteil der Robotik womöglich Rauschen misst.



Beide Ansätze durchlaufen dieselbe blinde, randomisierte Evaluation. Die Aussage des Papers ist nicht, dass Roboter-Foundation-Models Zero-Shot-Generalisten sind, sondern dass Vortraining die Dateneffizienz und Robustheit verbessern kann, wenn man sorgfältig misst.

Was die Autoren getan haben

Sie bauten LBMs in einem bestimmten, konkreten Sinn: **diffusionsbasierte visuomotorische Policies** (ein Diffusion Transformer, der Kamerabilder, eine kurze Sprachanweisung und die eigenen Gelenkstellungen des Roboters liest und in kurzen Schüben Motorbefehle mit 10 Hz ausgibt). Diese wurden **auf rund 1.700 Stunden** Roboterdemonstrationen **vortrainiert** — über 500 verschiedene, hausintern gesammelte Aufgaben plus öffentliche Datensätze — und dann **auf einzelne Aufgaben feinabgestimmt** (finetuned). Der Vergleich läuft durchweg gegen eine **von Grund auf trainierte Einzelaufgaben-Policy**, die nur auf den Daten dieser einen Aufgabe beruht.

Das Herzstück des Papers ist jedoch die *Evaluation*, die die Autoren als eigentliches Ergebnis behandeln. Um sich nicht selbst zu täuschen, verwendeten sie:

- **Blinde, randomisierte A/B-Tests** in der realen Welt — die Person, die den Roboter bediente, wusste nicht, welche Policy gerade getestet wurde, und die Reihenfolge war zufällig.
- **Kontrollierte, wiederholbare Ausgangsbedingungen** — die Operatoren glichen die Szene vor jedem Durchgang an eine Bildüberlagerung an.
- **Hohe Versuchszahlen** — 50 reale Durchläufe pro Aufgabe, pro Policy, pro Bedingung; 200 pro Aufgabe in der Simulation. Insgesamt: etwa **1.800 blinde reale Durchläufe und mehr als 47.000 Simulationsdurchläufe**.
- **Saubere Statistik** — Bayes'sche Schätzungen der Erfolgswahrscheinlichkeit und paarweise Hypothesentests mit Korrektur für multiple Vergleiche, statt überlappende Fehlerbalken nach Augenmaß zu deuten. Sie ließen sogar einen Qualitätssicherungs-Durchgang über ein Viertel der von Menschen bewerteten Versuche laufen, um den Bewertungsfehler zu messen.

[video pending]

Direkter Vergleich von Modellen beim Decken eines Frühstückstisches: (links) Einzelaufgaben-Basislinie und (rechts) LBM. Beide Videos laufen in 1-facher Geschwindigkeit. Dies ist eine evaluierte Aufgabe, kein Beleg für universelle Autonomie.

Diese Maschinerie ist der Punkt. Das ganze Paper ist ein Argument dafür, dass man ohne sie eine echte Verbesserung nicht von Glück unterscheiden kann.

Was sie fanden

Feinabgestimmte Großmodelle schlagen von Grund auf trainierte Einzelaufgaben-Modelle — im Durchschnitt. Über die Aufgaben hinweg aggregiert, übertraf ein LBM, das vortrainiert und dann auf eine Aufgabe feinabgestimmt wurde, verlässlich eine von Grund auf trainierte Policy für dieselbe Aufgabe, in Simulation wie in der realen Welt, und die Trennung war statistisch signifikant. Auf einzelnen Aufgaben war das feinabgestimmte LBM in fast jedem Fall statistisch gleich gut oder besser als das von Grund auf trainierte (3/3 reale Aufgaben, 15/16 Simulationsaufgaben).

Der größte, klarste Gewinn ist die Dateneffizienz. Ein feinabgestimmtes LBM erreichte die Leistung

des von Grund auf trainierten mit rund **3–5× weniger aufgabenspezifischen Daten**. Bei einer realen Aufgabe (einen Frühstückstisch decken) schlug ein LBM, das auf nur **15 %** der Demonstrationen feinabgestimmt war, eine von Grund auf trainierte Policy, die auf **100 %** davon trainiert war.

Vortraining hilft am meisten, wenn sich die Bedingungen verschieben. Wurde die Testumgebung absichtlich von den Trainingsbedingungen weg gestört („Distribution Shift"), *wuchs* der Vorteil des feinabgestimmten LBM. In einem Simulationssatz schlug es das von Grund auf trainierte unter normalen Bedingungen statistisch auf 3 von 16 Aufgaben, unter Distribution Shift aber auf 10 von 16. Da reale Einsätze immer von den Trainingsbedingungen abweichen, ist diese Robustheit wohl der praktisch wichtigste Befund.

Mehr Vortrainingsdaten halfen, gleichmäßig. Die Leistung stieg stetig, während sie Vortrainingsdaten hinzufügten — mit **keinem plötzlichen Sprung oder „emergenten" Satz** bei den getesteten Größenordnungen. Nützlich, vorhersehbar, undramatisch.

Aber die Generalist-ohne-Finetuning-Geschichte hielt nicht stand. Ein vortrainiertes LBM, *zero-shot* eingesetzt — ohne aufgabenspezifisches Finetuning —, schlug Einzelaufgaben-Policies *nicht* konsistent. Ein einzelnes Netz konnte viele Aufgaben zugleich, aber der „einfach prompten"-Traum bestätigte sich hier nicht; die Autoren führen einen Teil davon auf die Fragilität ihres kleinen Sprach-Encoders zurück.

Und die Zugewinne waren klein genug, um leicht übersehen — oder vorgetäuscht — zu werden. Viele der Effekte wurden erst mit den größer-als-üblichen Stichproben und sorgfältigen Tests sichtbar. Die Autoren stellen klar fest, dass angesichts der Effektgrößen und des Rauschens **ein erhebliches Risiko besteht, dass viele Robotik-Paper statistisches Rauschen messen.** Sie fanden auch, dass eine banale Entscheidung — wie die Daten *normalisiert* werden — die Ergebnisse stärker beeinflusste als architektonische Änderungen, und dass ein Normalisierungs-**Bug** im Vortraining erst *nach* Abschluss der Evaluationen zutage trat.

Was das wahrscheinlich bedeutet

Die vertretbare Lesart: großangelegtes Vortraining auf vielfältigen Roboterdaten ist ein echter, lohnender Baustein — es senkt den Datenbedarf pro neuer Aufgabe und macht Policies robuster, wenn die Welt nicht dem Training entspricht. Das stützt tatsächlich die Richtung, auf die das Feld setzt. Aber die Gewinne sind *bescheiden und bedingt* (sie zeigen sich meist erst nach dem Finetuning und sind am klarsten im Aggregat und unter Belastung), nicht die Ankunft eines einsatzfertigen Universalroboters.

Die stillere, wichtigere Bedeutung ist methodisch. Das Paper ist im Grunde ein Messstab: Es zeigt, wie viel Evidenz es tatsächlich braucht, um eine vertrauenswürdige Aussage über eine Roboter-Policy zu treffen, und legt nahe, dass ein Großteil der veröffentlichten Begeisterung auf zu wenig ruht. Das ist ein Korrektiv, das das Feld nötiger braucht als ein weiteres Modell.

Was das *nicht* beweist

- Es ist **kein Universalroboter.** Die Gewinne sind für eine bestimmte Architektur (Diffusion-Policies) belegt, pro Aufgabe feinabgestimmt, in kontrollierten Umgebungen, aus teleoperierten Demonstrationen — kein autonomer Roboter, der beliebige neue Aufgaben auf Kommando erledigt.
- Es bestätigt **nicht** den **Zero-Shot**-Einsatz. Ohne Finetuning schlug das Großmodell die Einzelaufgaben-Basislinien nicht konsistent.
- Es ist **kein** Beleg für einen **„emergenten Sprung".** Skalierung verbesserte die Dinge gleichmäßig; hier gibt es keine Unstetigkeit, die „und dann wurde es plötzlich fähig"-Erzählungen stützt.
- Die Zahlen sind **relativ und laborgebunden.** Die absoluten Erfolgsraten wurden absichtlich auf ~50 % eingestellt, um Vergleiche empfindlich zu machen; sie sind kein Maß für reale Zuverlässigkeit, und die Arbeit ist eine Architektur aus einem Labor.
- Es klärt **nicht**, **warum** eine Policy gelingt oder scheitert, und mehrere konkrete Aufgaben, bei denen das Großmodell *schlechter* war, werden berichtet, aber nicht erklärt.
- Es sagt **nichts über Sicherheit, Autonomie oder Einsatz** außerhalb des Evaluationsaufbaus.

Wie belastbar sind die Belege?

Für die zentralen Vergleichsaussagen — feinabgestimmte LBMs schlagen von Grund auf trainierte Basislinien im Aggregat, brauchen mehrfach weniger Daten und sind robuster unter Distribution Shift — ist die Evidenz stark *und* ungewöhnlich gut kontrolliert: blind, randomisiert, große Stichprobe, statistisch getestet, mit einem QA-Durchgang über die Bewertung. Das ist der seltene Fall, in dem die Methodik solide genug ist, um die Kernschlüsse fast für bare Münze zu nehmen.

Die ehrlichen Vorbehalte bringen die Autoren selbst vor. Ihre Fehlerbalken erfassen die Zufälligkeit der *Evaluation*, aber **nicht** die des *Trainings* — trainiert man dasselbe Modell zweimal, könnte man eine merklich andere Policy erhalten, und diese Variation steckt nicht in der Statistik. Reale Aufgaben hatten je 50 Versuche, genug, um mittlere Effekte zu erwischen, aber anfällig dafür, kleine zu verfehlen. Die Sprachkonditionierung nutzte einen bescheidenen Encoder, sodass Aussagen über „dem Roboter einfach sagen, was er tun soll" bei größeren Systemen anders ausfallen könnten. Und da ist die offene Nennung eines nachträglich entdeckten Normalisierungs-Bugs. Nichts davon versenkt die Hauptbefunde, aber es ist genau die Art von Dingen, die das Feld laut Paper üblicherweise beiseiteschiebt.

Eine Quellen-Anmerkung, im selben Geist: Dieser Erklärtext beruht auf dem Preprint der Autoren. Wir konnten die in der Zeitschrift veröffentlichte Fassung nicht abrufen und haben daher nicht auf Änderungen zwischen Preprint und veröffentlichtem Text geprüft.

Warum das wichtig ist

„Roboter-Foundation-Models" ist ein Ausdruck wie geschaffen fürs Überziehen, und eine Studie wie diese lässt sich leicht in beide Richtungen fehldeuten — als triumphales *„es funktioniert!"* oder als abwiegelndes *„alles Hype"*. Die zutreffende Deutung ist nützlicher als beide: Vortraining auf vielfältigen Daten liefert echte, messbare, aber moderate Vorteile — vor allem weniger Daten pro Aufgabe und mehr Robustheit —, und der Weg verbessert sich mit Skalierung vorhersehbar.

Der tiefere Grund, warum es zählt: Das Paper richtet seine Strenge auf das eigene Feld. Indem es zeigt, dass die echten Effekte klein genug sind, um unter schlampiger Evaluation zu verschwinden, und dass eine langweilige Entscheidung wie die Datennormalisierung eine clevere neue Architektur überwiegen kann, macht es geltend, dass ein Großteil des Fortschritts beim Roboter-Lernen eine solidere Messung braucht, bevor man ihm glauben darf. Ein Paper, das seine Glaubwürdigkeit darauf verwendet, den Unterschied zwischen einem Ergebnis und einem Wunsch zu überwachen, tut etwas Selteneres — und Wertvolleres — als eine Rangliste anzuführen.

Kurz zusammengefasst

Forschende am Toyota Research Institute trainierten „Large Behavior Models" — diffusionsbasierte Roboter-Policies, vortrainiert auf ~1.700 Stunden vielfältiger Manipulationsdaten — und testeten sie gegen von Grund auf trainierte Einzelaufgaben-Policies mit einem ungewöhnlich strengen Protokoll: blind, randomisiert, große Stichprobe (≈1.800 reale und über 47.000 Simulationsversuche), mit echter Statistik. Nach aufgabenweisem Finetuning schnitten die Großmodelle im Aggregat verlässlich besser ab, brauchten rund **3–5× weniger aufgabenspezifische Daten** und waren **robuster bei verschobenen Bedingungen**, wobei die Leistung mit wachsenden Vortrainingsdaten gleichmäßig stieg. Aber **ohne Finetuning** eingesetzt schlugen sie Einzelaufgaben-Modelle *nicht* konsistent, mehrere Effekte waren klein genug, dass erst die großen Stichproben sie sichtbar machten, und eine banale Datennormalisierungs-Entscheidung zählte mehr als die Architektur. Es ist solide, maßvolle Stütze für die Richtung der Roboter-Foundation-Models — **kein** Universalroboter, kein Zero-Shot-Generalist und kein „emergenter Sprung" — plus eine pointierte Warnung, dass ein Großteil der Robotik womöglich Rauschen misst.

Nüchtern geprüft

Was das Paper zeigt: Mit einer strengen, blinden, statistisch belastbaren Evaluation (≈1.800 reale + über 47.000 Simulationsdurchläufe) übertreffen multitask-vortrainierte-und-dann-feinabgestimmte Diffusion-Policies (LBMs) von Grund auf trainierte Einzelaufgaben-Policies im Aggregat, erreichen gleichwertige Leistung mit ~3–5× weniger aufgabenspezifischen Daten und sind robuster unter Distribution Shift; die Leistung skaliert gleichmäßig mit den Vortrainingsdaten.

Was plausibel, aber nicht bewiesen ist: Dass diese Vorteile sich auf viel größere Vision-Language-

Action-Modelle übertragen (ihr Sprach-Encoder war klein); dass die gleichmäßige Skalierung über den getesteten Datenbereich hinaus anhält.

Was es nicht zeigt: Einen Universal- oder Zero-Shot-Roboter (kein Finetuning → kein konsistenter Vorteil); irgendeinen „emergenten" Fähigkeitssprung; Erklärungen für konkrete Aufgaben-Fehlschläge; reale Zuverlässigkeit in absoluten Zahlen (die Erfolgsraten waren für die Empfindlichkeit nahe 50 % eingestellt); irgendetwas über Sicherheit oder autonomen Einsatz.

Wesentliche Einschränkungen: Die Statistik erfasst die Zufälligkeit der Evaluation, aber nicht die der Trainingsläufe; 50 reale Versuche pro Aufgabe könnten kleine Effekte verfehlen; eine Architektur und ein Labor; bescheidener Sprach-Encoder; ein Datennormalisierungs-Bug wurde nach den Evaluationen gefunden; Analyse auf Basis des Preprints (veröffentlichte Fassung nicht geprüft).

Wie viel Vertrauen sollte eine allgemeine Leserin haben? Hohes darin, dass Multitask-Vortraining plus Finetuning echte, moderate Vorteile bringt — besonders Dateneffizienz und Robustheit — und dass diese ungewöhnlich sorgfältig gemessen wurden. Hohes darin, dass dies **kein** Universal- oder Zero-Shot-Roboter und **kein** emergenter Sprung ist. Mittleres dazu, wie weit die Gewinne auf größere Modelle skalieren. Und ernst zu nehmen: die Warnung der Autoren selbst, dass die Effekte des Feldes klein genug sind, dass unterdimensionierte Studien Rauschen berichten könnten. Angemessene Haltung: maßvoller Optimismus gegenüber dem Ansatz und gesunde Skepsis gegenüber Roboter-KI-Ergebnissen, denen diese Art statistischer Absicherung fehlt.

Quellen

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/`