The Clean Paper

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

Apakah 'foundation model' robot besar benar-benar bekerja lebih baik? Jawaban hati-hati: sedikit ya, dan kebanyakan studi sulit mengetahuinya

Toyota Research Institute melatih 'large behavior models' — kebijakan robot yang dipretrain pada ~1.700 jam data manipulasi beragam — lalu mengujinya terhadap kebijakan satu tugas from-scratch dengan ketelitian tidak biasa: percobaan buta, acak, dan bersampel besar (~1.800 dunia nyata, 47.000+ simulasi) dengan statistik sungguhan. Setelah finetuning per tugas, model besar rata-rata lebih baik, membutuhkan kira-kira 3–5× lebih sedikit data khusus tugas, dan lebih tahan ketika kondisi berubah; performa naik halus dengan lebih banyak data pretraining. Namun tanpa finetuning mereka tidak secara konsisten mengalahkan model satu tugas, beberapa efek begitu kecil hingga hanya terlihat dengan sampel besar, dan pilihan normalisasi data biasa lebih penting daripada arsitektur. Ini dukungan terukur bagi arah foundation model robot — bukan robot serbaguna, bukan generalis zero-shot, bukan 'lompatan emergen' — sekaligus peringatan bahwa banyak penelitian robotika mungkin sedang mengukur derau.

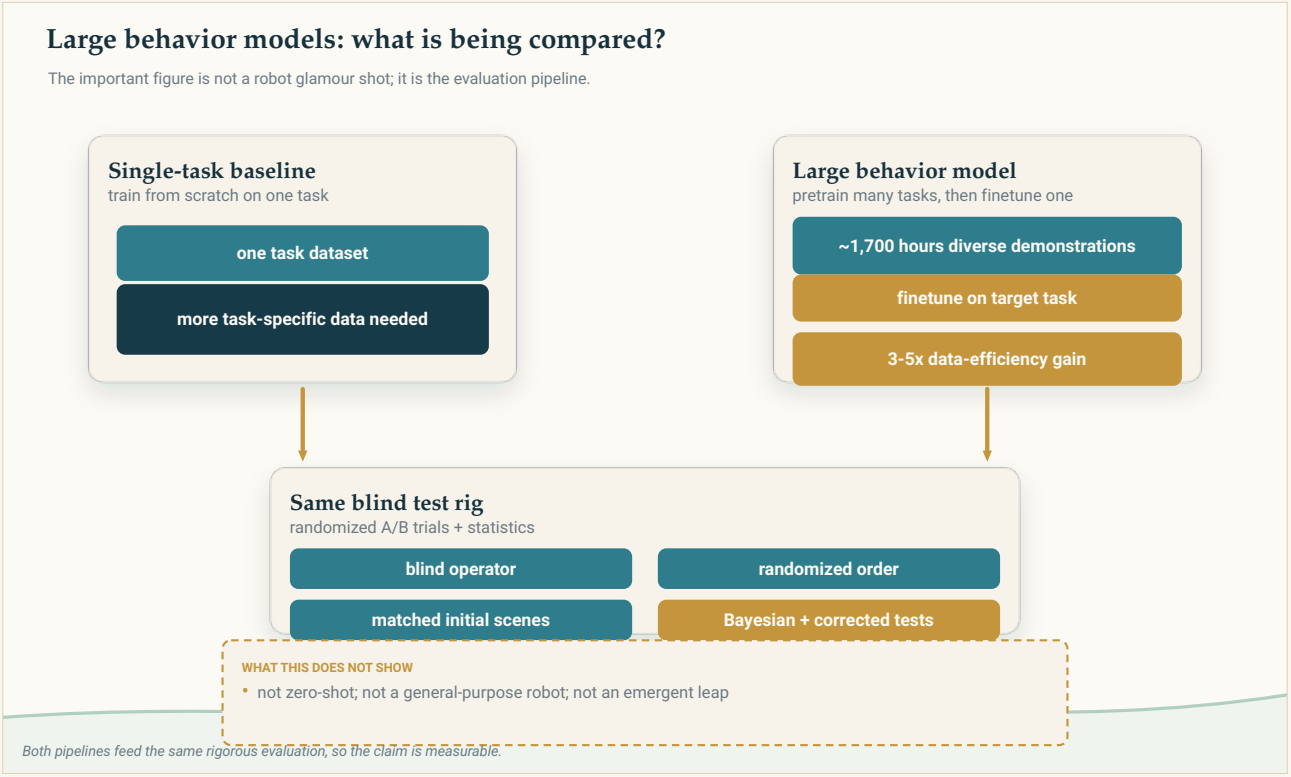
Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

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

Makalah robot yang sebenarnya berbicara tentang kejujuran

Robotika sedang dilanda demam **model fondasi** (*foundation model*). Gagasannya, dipinjam dari AI bahasa dan gambar, sangat menggoda: alih-alih melatih robot untuk satu tugas setiap kali, latih satu **large behavior model (LBM)** pada kumpulan demonstrasi yang sangat besar dan beragam, lalu gunakan prapelatihan itu agar sistem lebih mudah beradaptasi dengan tugas baru. Antusiasme — dan investasinya — sangat besar. Judul berita nyaris menulis dirinya sendiri: *otak robot serbaguna sudah tiba.*

Makalah dari Toyota Research Institute ini menarik justru karena menolak menulis judul tersebut. Kontribusi sebenarnya bukan robot yang lebih mencolok. Ia adalah pemeriksaan keras terhadap pertanyaan yang tampak sederhana tetapi menipu — *apakah pendekatan model besar benar-benar bekerja lebih baik, dan bagaimana kita bisa tahu?* — yang dijawab dengan kehati-hatian statistik yang, menurut para penulis sendiri, tidak biasa di bidang ini. Hasilnya adalah "ya, tetapi" yang benar-benar berguna, sekaligus peringatan bahwa banyak penelitian robotika mungkin sedang mengukur derau.



Kedua pendekatan masuk ke evaluasi buta dan acak yang sama. Klaim makalah bukan bahwa model fondasi robot sudah menjadi generalis *zero-shot*, melainkan bahwa **prapelatihan** dapat meningkatkan efisiensi data dan ketahanan jika manfaatnya diukur dengan hati-hati.

Original The Clean Paper diagram CC BY 4.0

Apa yang dilakukan para penulis

Mereka membangun LBM dalam arti yang spesifik dan konkret: **kebijakan visuomotor berbasis difusi** (Diffusion Transformer yang membaca gambar kamera, instruksi bahasa singkat, dan posisi sendi robot sendiri, lalu mengeluarkan rangkaian pendek perintah motor pada 10 Hz). Model-model ini menjalani **prapelatihan pada sekitar 1.700 jam** demonstrasi robot — lebih dari 500 tugas berbeda yang dikumpulkan sendiri, ditambah kumpulan data publik — lalu menjalani **fine-tuning** untuk tiap tugas. Sepanjang makalah, pembandingnya adalah **kebijakan satu tugas yang dilatih dari nol** hanya dengan data dari tugas tersebut.

Namun jantung makalah sebenarnya adalah *evaluasi*, yang diperlakukan penulis sebagai hasil utama. Untuk menghindari menipu diri sendiri, mereka menggunakan:

- **Uji A/B buta dan acak** di dunia nyata — operator yang menjalankan robot tidak tahu kebijakan mana yang sedang diuji, dan urutan pengujian diacak.
- **Kondisi awal yang terkontrol dan dapat diulang** — sebelum setiap percobaan, operator mencocokkan adegan dengan overlay gambar.
- **Jumlah percobaan besar** — 50 percobaan dunia nyata per tugas, per kebijakan, per kondisi; 200 per tugas dalam simulasi. Totalnya sekitar **1.800 percobaan dunia nyata yang buta dan lebih dari 47.000 percobaan simulasi**.
- **Statistik yang semestinya** — estimasi Bayesian terhadap probabilitas keberhasilan dan uji hipotesis berpasangan dengan koreksi perbandingan ganda, bukan sekadar menatap batang galat yang tumpang tindih. Mereka bahkan melakukan pemeriksaan jaminan mutu terhadap seperempat percobaan yang dinilai manusia untuk mengukur kesalahan penilaian.

[video pending]

Perbandingan berdampingan model yang menata meja sarapan: (kiri) model pembanding satu tugas, dan (kanan) LBM. Kedua video diputar pada kecepatan 1x. Ini satu tugas yang dievaluasi, bukan bukti otonomi serbaguna.

Perangkat pengukuran inilah inti ceritanya. Seluruh makalah pada dasarnya berargumen bahwa tanpa semua ini, kita tidak dapat membedakan peningkatan nyata dari keberuntungan.

Apa yang mereka temukan

Setelah fine-tuning, model besar rata-rata mengungguli model satu tugas yang dilatih dari awal. Jika hasil berbagai tugas digabungkan, LBM yang telah menjalani prapelatihan lalu fine-tuning secara andal mengungguli kebijakan yang dilatih dari awal pada data tugas yang sama, baik dalam simulasi maupun dunia nyata; perbedaannya signifikan secara statistik. Pada tingkat tugas individual, LBM setelah fine-tuning secara statistik sama baik atau lebih baik daripada model yang dilatih dari awal dalam hampir semua kasus (3/3 tugas dunia nyata, 15/16 tugas simulasi).

Kemenangan terbesar dan paling jelas adalah efisiensi data. LBM setelah fine-tuning mencapai

performa setara model yang dilatih dari awal dengan kira-kira **3–5× lebih sedikit data khusus tugas**. Pada satu tugas dunia nyata (menata meja sarapan), LBM yang menjalani fine-tuning hanya dengan **15%** demonstrasi mengalahkan kebijakan yang dilatih dari awal dengan **100%** demonstrasi.

Prapelatihan paling membantu ketika kondisi berubah. Ketika lingkungan pengujian sengaja digeser dari kondisi pelatihan ("pergeseran distribusi"), keunggulan LBM setelah fine-tuning justru *meningkat*. Dalam satu kelompok simulasi, model itu secara statistik mengungguli pembanding yang dilatih dari awal pada 3 dari 16 tugas dalam kondisi normal, tetapi pada 10 dari 16 tugas ketika distribusinya bergeser. Karena penerapan nyata selalu berbeda sedikit dari kondisi pelatihan, ketahanan ini mungkin merupakan hasil yang paling penting secara praktis.

Lebih banyak data prapelatihan memberi manfaat bertahap. Performa meningkat terus saat data prapelatihan ditambah — **tanpa lompatan mendadak atau "emergent"** pada skala yang diuji. Berguna, dapat diprediksi, tidak dramatis.

Namun klaim tentang generalis tanpa fine-tuning tidak bertahan. LBM yang hanya mengandalkan prapelatihan dan digunakan secara *zero-shot* — tanpa fine-tuning khusus tugas — *tidak* secara konsisten mengungguli kebijakan satu tugas. Satu jaringan memang dapat melakukan banyak tugas sekaligus, tetapi mimpi "cukup beri prompt" tidak terbukti di sini; para penulis menduga sebagian masalah berasal dari encoder bahasa mereka yang relatif kecil dan rapuh.

Dan peningkatannya cukup kecil sehingga mudah tidak terlihat — atau tampak palsu. Banyak efek baru terlihat dengan ukuran sampel yang lebih besar dari biasanya dan uji yang hati-hati. Para penulis mengatakan terus terang bahwa, dengan ukuran efek dan derau yang ada, **ada risiko signifikan bahwa banyak makalah robotika sebenarnya sedang mengukur derau statistik.** Mereka juga menemukan bahwa pilihan membosankan — bagaimana data *dinormalisasi* — memengaruhi hasil lebih besar daripada perubahan arsitektur, dan sebuah **kesalahan** normalisasi pada prapelatihan baru terungkap *setelah* evaluasi selesai.

Apa arti yang paling mungkin

Pembacaan yang paling masuk akal: prapelatihan skala besar pada data robot yang beragam adalah bahan nyata dan berguna — membuat setiap tugas baru memerlukan lebih sedikit data dan membuat kebijakan lebih tahan ketika dunia tidak persis sama dengan data pelatihan. Itu benar-benar mendukung arah yang sedang dipertaruhkan bidang ini. Namun peningkatannya *moderat dan bersyarat* (kebanyakan muncul setelah fine-tuning, dan paling jelas secara agregat serta saat kondisi diganggu), bukan kedatangan robot umum siap pakai.

Makna yang lebih tenang dan mungkin lebih penting bersifat metodologis. Makalah ini, pada dasarnya, adalah alat ukur: menunjukkan seberapa banyak bukti yang benar-benar dibutuhkan untuk membuat klaim tepercaya tentang kebijakan robot, dan menyiratkan bahwa cukup banyak kegembiraan yang diterbitkan mungkin bertumpu pada terlalu sedikit data. Bidang ini lebih membutuhkan koreksi seperti itu daripada satu model baru lagi.

Apa yang *tidak* dibuktikan

- Ini **bukan robot serbaguna**. Keunggulan ditunjukkan untuk arsitektur spesifik (kebijakan difusi) yang menjalani fine-tuning untuk tiap tugas, dalam kondisi terkontrol, dari demonstrasi teleoperasi — bukan robot otonom yang dapat melakukan pekerjaan baru sebarang berdasarkan perintah.
- Ini **tidak** memvalidasi penggunaan **zero-shot**. Tanpa fine-tuning, model besar tidak secara konsisten mengungguli model pembanding satu tugas.
- Ini **bukan** bukti **"lompatan emergen"**. Peningkatan skala memperbaiki performa secara bertahap; tidak ada diskontinuitas yang mendukung narasi "lalu tiba-tiba menjadi mampu".
- Angkanya **relatif dan terikat laboratorium**. Tingkat keberhasilan absolut sengaja disetel mendekati ~50% agar perbandingan sensitif; itu bukan ukuran reliabilitas dunia nyata, dan pekerjaan ini hanya satu arsitektur dari satu laboratorium.
- Ini **tidak** menjelaskan **mengapa** suatu kebijakan berhasil atau gagal, dan beberapa tugas spesifik di mana model besar justru *lebih buruk* dilaporkan tetapi tidak dijelaskan.
- Makalah ini **tidak mengatakan apa pun tentang keselamatan, otonomi, atau penerapan** di luar rig evaluasi.

Seberapa kuat buktinya?

Untuk klaim komparatif utama — LBM setelah fine-tuning mengungguli pembanding yang dilatih dari awal secara agregat, memerlukan beberapa kali lebih sedikit data, dan lebih tahan terhadap pergeseran distribusi — buktinya kuat *dan* tidak biasa karena sangat terkontrol: buta, acak, sampel besar, diuji statistik, dengan pemeriksaan QA pada penilaian. Ini kasus langka ketika metodologi cukup solid sehingga kesimpulan utama dapat diambil hampir apa adanya.

Keterbatasan jujurnya adalah hal-hal yang diangkat penulis sendiri. Batang galat mereka menangkap keacakan *evaluasi* tetapi **tidak** keacakan *pelatihan* — latih model yang sama dua kali dan kebijakan yang dihasilkan bisa berbeda bermakna, dan variasi itu tidak masuk statistik. Tugas dunia nyata memiliki 50 percobaan masing-masing, cukup untuk menangkap efek sedang tetapi bisa melewatkan efek kecil. Pengkondisian bahasa memakai encoder sederhana, jadi klaim tentang "cukup katakan kepada robot apa yang harus dilakukan" mungkin berbeda pada sistem lebih besar. Dan ada pengungkapan terbuka tentang kesalahan normalisasi yang ditemukan belakangan. Tidak satu pun menjatuhkan hasil utama, tetapi justru hal-hal semacam inilah yang menurut makalah sering dikesampingkan oleh bidang ini.

Satu catatan sumber dalam semangat yang sama: penjelasan ini didasarkan pada pracetak penulis. Kami tidak berhasil mendapatkan versi yang diterbitkan jurnal, sehingga belum memeriksa apakah ada perubahan antara pracetak dan teks terbit.

Mengapa ini penting

Istilah "model fondasi robot" sangat mudah memancing klaim berlebihan, dan studi seperti ini mudah disalahbaca ke dua arah — sebagai kemenangan *"ternyata berhasil!"* atau cibiran *"semuanya terlalu dibesar-besarkan."* Pembacaan yang akurat lebih berguna dari keduanya: prapelatihan pada data beragam memberi manfaat nyata, terukur, tetapi moderat — terutama lebih sedikit data per tugas dan lebih banyak ketahanan — dan jalurnya membaik secara dapat diprediksi bersama skala.

Alasan yang lebih dalam adalah makalah ini mengarahkan ketelitian tersebut pada bidangnya sendiri. Dengan menunjukkan bahwa efek asli cukup kecil untuk hilang jika evaluasinya ceroboh, dan bahwa pilihan membosankan seperti normalisasi data dapat lebih penting daripada arsitektur baru yang cerdik, makalah ini membuat argumen bahwa banyak kemajuan dalam pembelajaran robot membutuhkan pengukuran yang lebih kokoh sebelum layak dipercaya. Makalah yang menggunakan kredibilitasnya untuk menjaga batas antara hasil dan keinginan melakukan sesuatu yang lebih langka — dan lebih berharga — daripada sekadar menduduki puncak papan peringkat.

Singkatnya

Peneliti Toyota Research Institute melatih **large behavior models** — kebijakan robot berbasis difusi yang menjalani prapelatihan pada ~1.700 jam data manipulasi beragam — lalu mengujinya terhadap kebijakan satu tugas yang dilatih dari nol dengan protokol yang sangat ketat: buta, acak, sampel besar (≈1.800 percobaan dunia nyata dan 47.000+ simulasi), dengan statistik sungguhan. Setelah fine-tuning untuk tiap tugas, model besar secara andal lebih baik secara agregat, membutuhkan kira-kira **3–5× lebih sedikit data khusus tugas**, dan **lebih tahan ketika kondisi bergeser**, dengan performa meningkat halus seiring bertambahnya data prapelatihan. Namun jika digunakan **tanpa fine-tuning**, model tersebut *tidak* secara konsisten mengalahkan model satu tugas, beberapa efek terlalu kecil untuk terlihat tanpa sampel besar, dan pilihan normalisasi data yang biasa ternyata lebih penting daripada arsitektur. Ini dukungan yang solid dan terukur bagi arah model fondasi robot — **bukan** robot serbaguna, bukan generalis zero-shot, dan bukan "lompatan emergen" — plus peringatan tajam bahwa banyak penelitian robotika mungkin sedang mengukur derau.

Penilaian tanpa berlebihan

Apa yang ditunjukkan makalah: Dengan evaluasi buta, acak, dan memiliki daya statistik (≈1.800 percobaan dunia nyata + 47.000+ simulasi), kebijakan difusi yang menjalani prapelatihan multitugas lalu fine-tuning (LBM) mengungguli kebijakan satu tugas yang dilatih dari awal secara agregat, mencapai performa setara dengan ~3–5× lebih sedikit data khusus tugas, dan lebih tahan terhadap pergeseran distribusi; performa meningkat bertahap seiring bertambahnya data prapelatihan.

Apa yang masuk akal tetapi belum terbukti: Bahwa manfaat ini berpindah ke model visi–bahasa–aksi yang jauh lebih besar (encoder bahasa mereka kecil); bahwa peningkatan skala yang bertahap

terus berlanjut di luar rentang data yang diuji.

Apa yang tidak ditunjukkan: Robot serbaguna atau zero-shot (tanpa fine-tuning → tidak ada keunggulan konsisten); lompatan kemampuan "emergen"; penjelasan kegagalan tugas tertentu; reliabilitas dunia nyata dalam angka absolut (tingkat keberhasilan sengaja disetel dekat 50% demi sensitivitas); apa pun tentang keselamatan atau penerapan otonom.

Keterbatasan utama: Statistik menangkap keacakan evaluasi tetapi tidak variasi antarproses pelatihan; 50 percobaan dunia nyata per tugas mungkin melewatkan efek kecil; satu arsitektur dan satu laboratorium; encoder bahasa sederhana; kesalahan normalisasi data ditemukan setelah evaluasi; analisis didasarkan pada pracetak; versi terbit belum dapat kami periksa.

Seberapa besar keyakinan yang layak dimiliki pembaca umum? Keyakinan tinggi bahwa prapelatihan multitugas ditambah fine-tuning memberi manfaat nyata tetapi moderat — terutama efisiensi data dan ketahanan — dan bahwa manfaat tersebut diukur dengan ketelitian tidak biasa. Keyakinan tinggi bahwa ini **bukan** robot serbaguna atau zero-shot dan **bukan** lompatan emergen. Keyakinan sedang tentang seberapa jauh keuntungan meningkat pada model yang lebih besar. Dan peringatan para penulis layak dianggap serius: efek di bidang ini cukup kecil sehingga studi dengan daya statistik rendah dapat keliru melaporkan derau sebagai hasil. Sikap yang tepat: optimisme terukur tentang pendekatan, dan skeptisisme sehat terhadap hasil robot-AI yang tidak memiliki dukungan statistik seperti ini.

Sumber

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/