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

Adakah "model asas" robot besar benar-benar berfungsi lebih baik? Jawapan teliti — sedikit ya, dan kebanyakan kajian sukar membezakannya

Toyota Research Institute melatih "large behavior models" — dasar robot yang dipralatih pada ~1,700 jam data manipulasi pelbagai — dan mengujinya terhadap dasar satu tugas dari awal dengan ketelitian luar biasa: percubaan buta, rawak, bersampel besar (~1,800 dunia sebenar, 47,000+ simulasi) dengan statistik sebenar. Selepas penalaan halus setiap tugas, model besar lebih baik secara purata, memerlukan kira-kira 3–5× kurang data khusus tugas, dan lebih teguh apabila keadaan berubah; prestasi meningkat secara lancar dengan lebih banyak data pralatihan. Tetapi tanpa penalaan halus ia tidak secara konsisten mengatasi model satu tugas, beberapa kesan cukup kecil hingga memerlukan sampel besar untuk dilihat, dan pilihan normalisasi data biasa lebih penting daripada seni bina. Ini sokongan terukur bagi arah model asas robot — bukan robot serba guna, bukan generalis zero-shot, bukan "lonjakan emergen" — serta amaran tajam bahawa banyak robotik mungkin mengukur hingar.

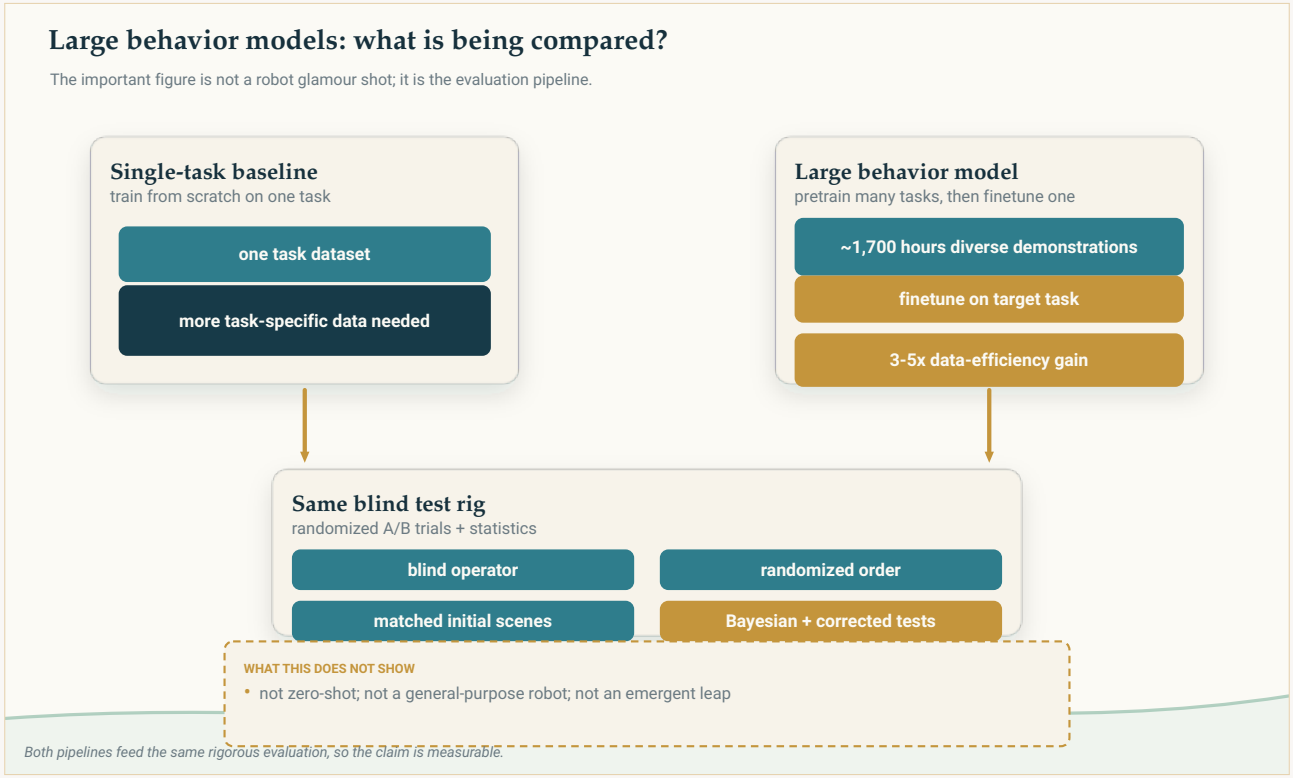
Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

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

Kertas robot yang subjek sebenarnya ialah kejujuran

Robotik sedang berada di tengah-tengah demam "model asas". Ideanya, dipinjam daripada AI bahasa dan imej, memang menggoda: daripada melatih robot satu tugas pada satu masa, latih satu **"large behavior model" (LBM)** besar pada timbunan demonstrasi yang sangat besar dan pelbagai, lalu dapatkan sistem yang berkebolehan luas dan cepat menyesuaikan diri. Keterujaan — dan pelaburan — sangat besar. Tajuk berita menulis dirinya sendiri: *otak robot serba guna sudah tiba.*

Kertas ini, daripada Toyota Research Institute, menarik tepat kerana ia enggan menulis tajuk itu. Sumbangan sebenarnya bukan robot yang lebih bergaya. Ia pemeriksaan keras terhadap soalan yang nampak mudah tetapi menipu — *adakah pendekatan model besar benar-benar berfungsi lebih baik, dan bagaimana kita boleh tahu?* — dijawab dengan tahap ketelitian statistik yang, menurut para penulis sendiri, luar biasa dalam bidang tersebut. Hasilnya ialah "ya, tetapi" yang benar-benar berguna, dan amaran bahawa sebahagian besar robotik mungkin sedang mengukur hingar.



Kedua-dua pendekatan masuk ke penilaian buta dan rawak yang sama. Dakwaan kertas bukan bahawa model asas robot ialah generalis zero-shot, tetapi bahawa pralatihan boleh meningkatkan kecekapan data dan keteguhan apabila diukur dengan teliti.

Original The Clean Paper diagram CC BY 4.0

Apa yang dilakukan oleh para penulis

Mereka membina LBM dalam erti yang khusus dan konkrit: **dasar visuomotor berasaskan difusi** (Diffusion Transformer yang membaca imej kamera, arahan bahasa pendek dan kedudukan sendi robot sendiri, kemudian menghasilkan kelompok pendek arahan motor pada 10 Hz). Sistem ini **dipralatih pada kira-kira 1,700 jam** demonstrasi robot — lebih 500 tugas berbeza yang dikumpul sendiri, ditambah set data awam — lalu **ditala halus** pada tugas individu. Perbandingan sepanjang kertas ialah terhadap **dasar satu tugas yang dilatih dari awal** menggunakan hanya data tugas tersebut.

Tetapi jantung kertas ialah *penilaian*, yang dianggap para penulis sebagai hasil utama. Untuk mengelakkan menipu diri sendiri mereka menggunakan:

- **Ujian A/B buta dan rawak** di dunia sebenar — manusia yang menjalankan robot tidak tahu dasar mana sedang diuji, dan urutannya dirawak.
- **Keadaan awal terkawal dan boleh diulang** — operator memadankan adegan kepada tindanan imej sebelum setiap percubaan.
- **Bilangan percubaan besar** — 50 pelaksanaan dunia sebenar bagi setiap tugas, setiap dasar, setiap keadaan; 200 bagi setiap tugas dalam simulasi. Jumlahnya sekitar **1,800 pelaksanaan buta di dunia sebenar dan lebih 47,000 pelaksanaan simulasi**.
- **Statistik yang betul** — anggaran Bayesian bagi kebarangkalian kejayaan, serta ujian hipotesis berpasangan dengan pembetulan perbandingan berganda, bukannya meneka daripada palang ralat yang bertindih. Mereka malah menjalankan pusingan jaminan kualiti pada satu perempat percubaan yang dinilai manusia untuk mengukur ralat pemarkahan.

[video pending]

Perbandingan bersebelahan model menyediakan meja sarapan: (kiri) garis dasar satu tugas, dan (kanan) LBM. Kedua-dua video dimainkan pada kelajuan 1×. Ini satu tugas yang dinilai, bukan bukti autonomi serba guna.

Mesin penilaian inilah intinya. Seluruh kertas ialah hujah bahawa tanpanya, anda tidak boleh membezakan peningkatan sebenar daripada nasib.

Apa yang mereka temui

Model besar yang ditala halus mengatasi model satu tugas dari awal — secara purata. Apabila digabungkan merentasi tugas, LBM yang dipralatih kemudian ditala halus pada satu tugas secara boleh dipercayai mengatasi dasar yang dilatih dari awal pada tugas sama, dalam simulasi dan dunia sebenar, dan perbezaannya signifikan secara statistik. Pada tugas individu, LBM ditala halus secara statistik sama baik atau lebih baik daripada model dari awal dalam hampir semua kes (3/3 tugas dunia sebenar, 15/16 tugas simulasi).

Kemenangan terbesar dan paling jelas ialah kecekapan data. LBM ditala halus mencapai prestasi setara model dari awal dengan kira-kira **3–5× kurang data khusus tugas**. Dalam satu tugas dunia

sebenar (menyediakan meja sarapan), LBM yang ditala halus menggunakan hanya **15%** demonstrasi mengatasi dasar dari awal yang dilatih pada **100%** demonstrasi.

Pralatihan paling membantu apabila keadaan berubah. Apabila persekitaran ujian sengaja diubah daripada keadaan latihan ("anjakan taburan"), kelebihan LBM ditala halus *bertambah*. Dalam satu set simulasi, ia mengatasi model dari awal secara statistik pada 3 daripada 16 tugas di bawah keadaan normal tetapi 10 daripada 16 di bawah anjakan taburan. Oleh sebab penggunaan dunia sebenar sentiasa menyimpang daripada keadaan latihan, keteguhan ini mungkin penemuan paling penting secara praktikal.

Lebih banyak data pralatihan membantu secara lancar. Prestasi meningkat secara berterusan apabila data pralatihan ditambah — dengan **tiada lonjakan mendadak atau "kemunculan" kemampuan** pada skala yang diuji. Berguna, boleh dijangka, tidak dramatik.

Tetapi cerita generalis tanpa penalaan halus tidak bertahan. LBM pralatih yang digunakan secara *zero-shot* — tanpa penalaan khusus tugas — tidak secara konsisten mengatasi dasar satu tugas. Satu rangkaian memang boleh melakukan banyak tugas sekali gus, tetapi impian "beritahu sahaja apa hendak dibuat" tidak disokong di sini; para penulis mengaitkan sebahagiannya dengan kerapuhan pengekod bahasa kecil mereka.

Dan peningkatannya cukup kecil untuk mudah terlepas — atau kelihatan palsu. Banyak kesan hanya muncul dengan saiz sampel yang lebih besar daripada biasa dan ujian yang teliti. Para penulis menyatakan secara terus bahawa, memandangkan saiz kesan dan hingar, **terdapat risiko besar bahawa banyak kertas robotik sebenarnya mengukur hingar statistik.** Mereka turut mendapati pilihan biasa — cara data *dinormalisasikan* — memberi kesan lebih besar daripada perubahan seni bina, dan satu **pepijat** normalisasi dalam pralatihan hanya ditemui *selepas* semua penilaian selesai.

Apa yang mungkin dimaksudkannya

Bacaan yang boleh dipertahankan: pralatihan berskala besar pada data robot yang pelbagai ialah bahan yang nyata dan berguna — ia mengurangkan data yang diperlukan bagi setiap tugas baharu dan menjadikan dasar lebih teguh apabila dunia tidak sepadan dengan latihan. Itu benar-benar menyokong arah yang sedang dipertaruhkan oleh bidang ini. Tetapi keuntungannya *sederhana dan bersyarat* (kebanyakannya muncul selepas penalaan halus, dan paling jelas apabila digabungkan serta di bawah tekanan), bukan ketibaan robot umum yang boleh dipasang terus.

Makna yang lebih senyap dan mungkin lebih penting ialah metodologi. Kertas ini pada dasarnya ialah pembaris pengukur: ia menunjukkan berapa banyak bukti yang sebenarnya diperlukan untuk membuat dakwaan boleh dipercayai tentang dasar robot, dan mencadangkan bahawa banyak keterujaan diterbitkan atas terlalu sedikit data. Bidang ini lebih memerlukan pembetulan seperti itu daripada satu lagi model.

Apa yang *tidak* dibuktikan oleh kajian ini

- Ia **bukan robot serba guna**. Kemenangan ditunjukkan untuk seni bina tertentu (dasar difusi) yang ditala halus bagi setiap tugas, dalam keadaan terkawal, daripada demonstrasi teleoperasi — bukan robot autonomi yang melakukan kerja baharu sewenang-wenangnya atas arahan.
- Ia **tidak** mengesahkan penggunaan **zero-shot**. Tanpa penalaan halus, model besar tidak secara konsisten mengatasi garis dasar satu tugas.
- Ia **bukan** bukti **"lonjakan emergen."** Penskalaan meningkatkan prestasi secara lancar; tiada ketakselanjaran di sini untuk menyokong cerita "kemudian ia tiba-tiba menjadi berkebolehan."
- Nombornya **relatif dan terikat makmal.** Kadar kejayaan mutlak sengaja ditala sekitar ~50% untuk menjadikan perbandingan sensitif; ia bukan ukuran kebolehpercayaan dunia sebenar, dan kerja ini ialah satu seni bina dari satu makmal.
- Ia **tidak** menyelesaikan **mengapa** sesuatu dasar berjaya atau gagal, dan beberapa tugas khusus di mana model besar berprestasi *lebih buruk* dilaporkan tetapi tidak diterangkan.
- Ia tidak mengatakan **apa-apa tentang keselamatan, autonomi atau penggunaan** di luar rig penilaian.

Sejauh mana kukuh buktinya?

Bagi dakwaan perbandingan teras — LBM ditala halus mengatasi garis dasar dari awal secara agregat, memerlukan beberapa kali kurang data, dan lebih teguh di bawah anjakan taburan — buktinya kuat *dan* luar biasa terkawal: buta, rawak, sampel besar, diuji secara statistik, dengan pusingan QA pada pemarkahan. Ini kes jarang metodologi cukup kukuh untuk mengambil kesimpulan utama hampir pada nilai muka.

Peringatan penting ialah yang dibangkitkan sendiri oleh para penulis. Palang ralat mereka menangkap rawak daripada *penilaian* tetapi **bukan** rawak daripada *latihan* — latih model yang sama dua kali dan anda mungkin mendapat dasar yang berbeza secara bermakna, dan variasi itu tidak berada dalam statistik. Tugas dunia sebenar mempunyai 50 percubaan setiap satu, cukup untuk menangkap kesan sederhana tetapi mungkin terlepas kesan kecil. Pengkondisian bahasa menggunakan pengekod sederhana, jadi dakwaan tentang "beritahu sahaja robot apa hendak dibuat" mungkin berbeza untuk sistem lebih besar. Dan ada pendedahan terbuka tentang pepijat normalisasi yang ditemui selepas penilaian. Tiada satu pun membatalkan hasil utama, tetapi semuanya tepat jenis perkara yang menurut kertas ini sering disapu ke tepi oleh bidang tersebut.

Satu nota sumber, dalam semangat yang sama: penerangan ini berasaskan pracetak para penulis. Kami tidak dapat mendapatkan versi yang diterbitkan jurnal, jadi kami belum menyemak sama ada terdapat perubahan antara pracetak dan teks terbitan.

Mengapa ia penting

"Model asas robot" ialah frasa yang dibuat untuk dakwaan berlebihan, dan kajian seperti ini mudah disalah baca ke dua arah — sebagai kemenangan *"ia berfungsi!"* atau penolakan *"ia berlebihan."* Bacaan tepat lebih berguna daripada kedua-duanya: pralatihan pada data pelbagai memberikan manfaat yang nyata, terukur tetapi sederhana — terutamanya kurang data setiap tugas dan lebih keteguhan — dan laluan bertambah baik secara boleh diramal dengan skala.

Sebab lebih mendalam ia penting ialah kertas ini mengarahkan ketegasannya kepada bidang sendiri. Dengan menunjukkan bahawa kesan sebenar cukup kecil untuk hilang di bawah penilaian cuai, dan bahawa pilihan membosankan seperti normalisasi data boleh melebihi seni bina baharu yang pintar, ia membina kes bahawa banyak kemajuan pembelajaran robot memerlukan pengukuran lebih kukuh sebelum boleh dipercayai. Kertas yang menggunakan kredibilitinya untuk mengawal sempadan antara hasil dan harapan melakukan sesuatu yang lebih jarang, dan lebih bernilai, daripada memuncaki papan pendahulu.

Ringkasan utama

Penyelidik di Toyota Research Institute melatih "large behavior models" — dasar robot berasaskan difusi yang dipralatih pada ~1,700 jam data manipulasi pelbagai — dan mengujinya terhadap dasar satu tugas yang dilatih dari awal dengan protokol luar biasa ketat: buta, rawak, sampel besar (≈1,800 percubaan dunia sebenar dan 47,000+ simulasi), dengan statistik sebenar. Selepas penalaan halus setiap tugas, model besar secara boleh dipercayai lebih baik secara agregat, memerlukan kira-kira **3–5× kurang data khusus tugas**, dan **lebih teguh apabila keadaan berubah**, dengan prestasi meningkat secara lancar apabila data pralatihan bertambah. Tetapi apabila digunakan **tanpa penalaan halus** ia tidak secara konsisten mengatasi model satu tugas, beberapa kesan begitu kecil sehingga hanya saiz sampel besar mendedahkannya, dan pilihan normalisasi data biasa lebih penting daripada seni bina. Ini sokongan yang kukuh dan terukur bagi arah model asas robot — **bukan** robot serba guna, bukan generalis zero-shot, dan bukan "lonjakan emergen" — serta amaran tajam bahawa sebahagian robotik mungkin sedang mengukur hingar.

Semakan tanpa gembar-gembur

Apa yang ditunjukkan oleh kertas ini: Dengan penilaian ketat, buta dan berkuasa statistik (≈1,800 pelaksanaan dunia sebenar + 47,000+ simulasi), dasar difusi yang dipralatih secara multitugas kemudian ditala halus (LBM) mengatasi dasar satu tugas dari awal secara agregat, mencapai prestasi setara dengan ~3–5× kurang data khusus tugas, lebih teguh di bawah anjakan taburan, dan prestasi meningkat secara lancar dengan data pralatihan.

Apa yang munasabah tetapi belum dibuktikan: Bahawa manfaat ini berpindah kepada model vision-language-action yang jauh lebih besar (pengekod bahasa mereka kecil); bahawa penskalaan lancar

berterusan melampaui julat data yang diuji.

Apa yang tidak ditunjukkannya: Robot serba guna atau zero-shot (tanpa penalaan halus → tiada kelebihan konsisten); sebarang lonjakan kemampuan "emergen"; penjelasan bagi kegagalan khusus tugas; kebolehpercayaan dunia sebenar secara mutlak (kadar kejayaan ditala sekitar 50% untuk sensitiviti); apa-apa tentang keselamatan atau penggunaan autonomi.

Had utama: Statistik menangkap rawak penilaian tetapi bukan rawak larian latihan; 50 percubaan dunia sebenar setiap tugas mungkin terlepas kesan kecil; satu seni bina dan satu makmal; pengekod bahasa sederhana; pepijat normalisasi data ditemui selepas penilaian; analisis berdasarkan pracetak (versi terbitan tidak disemak).

Sejauh mana pembaca umum patut yakin? Tinggi bahawa pralatihan multitugas ditambah penalaan halus memberi manfaat sebenar tetapi sederhana — khususnya kecekapan data dan keteguhan — dan bahawa ia diukur dengan luar biasa teliti. Tinggi bahawa ini **bukan** robot serba guna atau zero-shot dan **bukan** lonjakan emergen. Sederhana tentang sejauh mana keuntungan itu berskala kepada model lebih besar. Dan patut diambil serius: amaran penulis sendiri bahawa kesan bidang ini cukup kecil sehingga kajian dengan kuasa statistik rendah mungkin melaporkan hingar. Sikap sesuai: optimisme terukur tentang pendekatan, dan skeptisisme sihat terhadap hasil AI robot yang tidak mempunyai sokongan statistik sekuat 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/