The CLEAN PAPER

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

ትልቅ የሮቦት behavior models በጥንቃቄ ሲፈተኑ፡- እውነተኛ ጥቅም አለ፣ ግን አጠቃላይ ሮቦት ገና የለም

Toyota Research Institute በ~1,700 ሰዓት የሮቦት ዳታ ላይ ቀድሞ የተሰለጠኑ large behavior models'ን በblind፣ randomised እና ትልቅ-sample evaluation ፈተነ። Finetune ከተደረጉ በኋላ ~3–5× ያነሰ task-specific data ፈለጉ እና ሁኔታ ሲለወጥ ይበልጥ ጽኑ ነበሩ። Zero-shot ጥቅም ግን በተደጋጋሚ አልታየም። ይህ መጠነኛ ድጋፍ ለrobot foundation models እና በመስኩ ደካማ ስታቲስቲክ ላይ ጠንካራ ማስጠንቀቂያ ነው።

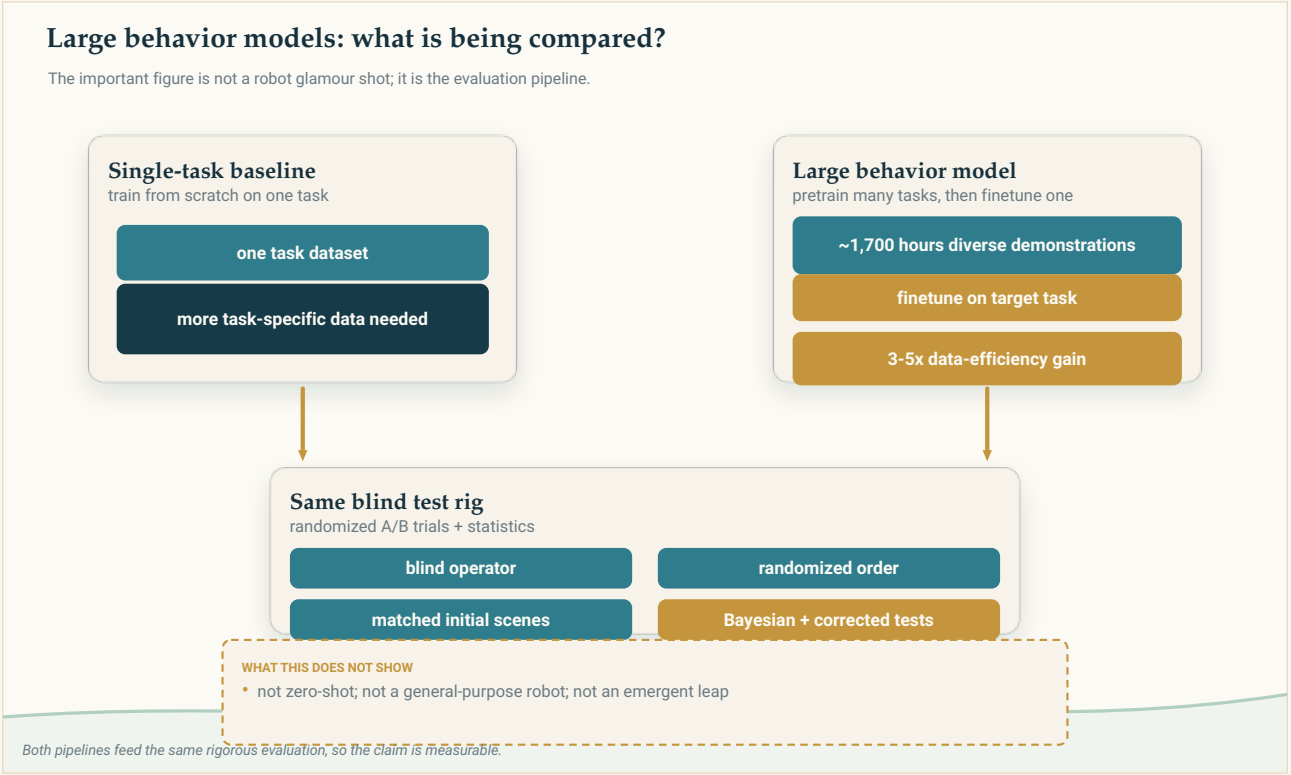
Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

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

ስለ ሮቦት የተጻፈ ጥናት፣ እውነተኛው ርዕሱ ግን ታማኝ መለኪያ ነው

Robotics አሁን “foundation ሞዴል” የሚባል ትልቅ ፍጥነት ውስጥ ነው። ከቋንቋና ከምስል AI የተወሰደው ሀሳብ ቀላልና ማራኪ ነው፤ ሮቦትን ለእያንዳንዱ ሥራ ለብቻው ከማሰልጠን ይልቅ አንድ ትልቅ large ባህሪ ሞዴል (LBM) በብዙና የተለያዩ ማሳያዎች ላይ አስቀድሞ ማሰልጠን፣ ከዚያም ብዙ ሥራዎችን ሊያደርግና ፈጥኖ ሊላመድ የሚችል ስርዓት ማግኘት። ጉጉቱና ኢንቨስትመንቱ ትልቅ ነው፣ እና “አጠቃላይ-ዓላማ የሮቦት አንጎሎች መጥተዋል” የሚል ርዕስ በራሱ ይጻፋል።

ከToyota Research Institute የመጣው ይህ ጥናት አስደሳች የሚሆነው ያንን ርዕስ ለመጻፍ ስለማይቸኩል ነው። ዋና አስተዋፅኦው ይበልጥ የሚያስደንቅ ሮቦት አይደለም። ቀላል የሚመስል ግን አስቸጋሪ ጥያቄን በጥንቃቄ ይመለከታል፤ ትልቅ ሞዴል በእውነት የተሻለ ይሠራል? እና እንደሚሠራ እንዴት እናውቃለን? ደራሲዎቹ በመስኩ ያልተለመደ የስታቲስቲክ ጥንቃቄ ተጠቅመው መልሰዋል። ውጤቱ “አዎን፣ ግን...” የሚል ጠቃሚ መልስ ነው፣ እና ብዙ የrobotics ጥናቶች እውነተኛ ውጤት ሳይሆን የስታቲስቲክ ጫጫታ ሊለኩ እንደሚችሉ ማስጠንቀቂያ ነው።



ሁለቱም አቀራረቦች በተመሳሳይ blind እና በዘፈቀደ የተመደበ ግምገማ ይፈተናሉ። ጥናቱ የሚለው robot foundation ሞዴሎች zero-shot አጠቃላይ ችሎታ አላቸው አይደለም፤ በጥንቃቄ ሲለኩ ቅድመ-ስልጠና የዳታ ውጤታማነትንና ጽናትን ሊያሻሽል እንደሚችል ነው።

Original The Clean Paper diagram CC BY 4.0

ደራሲዎቹ ምን አደረጉ?

LBMን በግልጽ ቴክኒካዊ ትርጉም ገነቡ፤ **ዲፊውዥን-based visuomotor policies**። አንድ ዲፊውዥን Transformer ከካሜራ ምስሎች፣ ከአጭር የቋንቋ መመሪያ እና ከሮቦቱ የመገጣጠሚያ አቀማመጥ መረጃ ይቀበላል፣ ከዚያም በ10 Hz አጭር የሞተር ትዕዛዞችን ያወጣል። ሞዴሎቹ **በ1,700 ሰዓት ገደማ** የሮቦት ማሳያዎች — በቤት ውስጥ የተሰበሰቡ ከ500 በላይ የተለያዩ ሥራዎች እና የህዝብ የመረጃ ስብስቦች — ላይ ቀድሞ ተሰልጥነው፣ ከዚያ ለእያንዳንዱ ሥራ finetune ተደረጉ። በሁሉም ንጽጽር ላይ መሠረታዊው ተፎካካሪ በዚያ አንድ ሥራ ዳታ ላይ ከዜሮ የተሰለጠነ single-**ተግባር** policy ነው።

የጥናቱ ልብ ግን ግምገማ ነው። ደራሲዎቹ እራሳቸውን እንዳያታልሉ የሚከተሉትን ተጠቀሙ፡-

- በእውነተኛው ዓለም blind, **በዘፈቀደ የተመደበ** A/B testing፤ ሮቦቱን የሚያስኬደው ሰው የትኛው policy እንደሚፈተን አያውቅም፣ የፈተናውም ቅደም ተከተል በየዘፈቀደ ነው።
- **ቁጥጥር ያለውና ሊደገም የሚችል መነሻ ሁኔታ**፤ ከእያንዳንዱ ሙከራ በፊት ኦፕሬተሮች ትዕይንቱን ከimage overlay ጋር ያዛምዱታል።
- **ብዙ የሙከራ ቁጥሮች**፤ ለእያንዳንዱ ሥራ፣ policy እና condition 50 የበእውነተኛ ዓለም rollouts፣ በማስመሰል ደግሞ 200። በጠቅላላው ወደ **1,800 blind በእውነተኛ ዓለም rollouts እና ከ47,000 በላይ ማስመሰል rollouts**።
- **ትክክለኛ ስታቲስቲክ**፤ የBayesian የsuccess probability ግምቶች፣ እና ለmultiple comparisons እርማት ያላቸው pairwise hypothesis ሙከራዎች። በሰው የተመዘኑት ሙከራዎች አንድ አራተኛ ላይ የQA ግምገማም አደረጉ፣ የscoring ስህተትን ለመለካት።

[video pending]

የቁርስ ጠረጴዛ የሚያዘጋጁ ሁለት ሞዴሎች ጎን-ለጎን ንጽጽር፤ በግራ single-ተግባር baseline፣ በቀኝ LBM። ሁለቱም ቪዲዮዎች በ1x ፍጥነት ናቸው። ይህ አንድ የተፈተነ ሥራ ነው፣ የአጠቃላይ-ዓላማ autonomy ማስረጃ አይደለም።

ይህ የመለኪያ ሂደት ራሱ ዋናው ነጥብ ነው። የጥናቱ ክርክር፤ ይህን ያህል ጥንቃቄ ካልነበረ እውነተኛ ማሻሻያን ከዕድል መለየት አይቻልም።

ምን አገኙ?

Finetune የተደረጉ ትልቅ ሞዴሎች በአማካይ ከfrom-scratch single-ተግባር ሞዴሎች በላይ ነበሩ። ብዙ ሥራዎችን በአንድ ላይ ሲያዩ፣ ቀድሞ የተሰለጠነና ከዚያ ለሥራው finetune የተደረገ LBM በማስመሰል እና በእውነተኛ world ውስጥ በዚያ ሥራ ዳታ ብቻ ከዜሮ የተሰለጠነ policyን በተደጋጋሚ አሸነፈ፣ ልዩነቱም ስታቲስቲካዊ ጉልህ ነበር። በእያንዳንዱ ሥራ ሲታይ የfinetuned LBM ከfrom-scratch ጋር በስታቲስቲክ እኩል ወይም የተሻለ ነበር በአብዛኛው፤ 3/3 በእውነተኛ ዓለም ሥራዎች፣ 15/16 ማስመሰል ሥራዎች።

ትልቁና ግልጹ ጥቅም የዳታ ውጤታማነት ነው። Finetuned LBM ከfrom-scratch ሞዴል ጋር እኩል የሆነ አፈጻጸም ለመድረስ በግምት **3–5× ያነሰ ተግባር-specific መረጃ** ብቻ ፈለገ። በአንድ በእውነተኛ ዓለም ሥራ — የቁርስ ጠረጴዛ ማዘጋጀት — በማሳያዎቹ **15%** ብቻ finetune የተደረገ LBM፣ **በ100%** ዳታ የተሰለጠነ from-scratch policyን አሸነፈ።

ሁኔታዎች ሲለወጡ pretraining ይበልጥ ይረዳል። የፈተና አካባቢውን ከስልጠና conditions በፍላጎት ሲያርቁት — **ስርጭት shift** — የfinetuned LBM ጥቅም ጨመረ። በአንድ ማስመሰል ስብስብ ውስጥ፣ በመደበኛ ሁኔታ ከ16

ሥራዎች 3 ላይ ከfrom-scratch የተሻለ ነበር፤ በስርጭት shift ግን 10/16 ላይ። በእውነተኛ አጠቃቀም ሁኔታዎች ከስልጠና ሁኔታ ሁልጊዜ ትንሽ ስለሚለዩ፣ ይህ ጽናት በተግባር ምናልባት በጣም አስፈላጊው ውጤት ነው።

ተጨማሪ pretraining መረጃ በቀጣይነት ረዳ። ዳታ ሲጨምሩ አፈጻጸሙ ቀስ በቀስ ጨመረ። በተፈተነው ስኬል ድንገተኛ "emergent" ዝላይ አልታየም። ጠቃሚ፣ ሊተነበይ የሚችል፣ ያልተደራማ ውጤት።

ግን ያለ finetuning የአጠቃላይ-ችሎታ ታሪኩ አልተደገፈም። Pretrained LBM *zero-shot* — ምንም ተግባር-specific finetuning ሳይደረግ — single-ተግባር policiesን በተደጋጋሚ አላሸነፈም። አንድ network ብዙ ሥራዎችን በአንድ ጊዜ ሊያደርግ ቢችልም "በቃ ትዕዛዝ ስጠው" የሚለው ህልም እዚህ አልተረጋገጠም። ደራሲዎቹ ከምክንያቶቹ አንዱ ትንሹ ቋንቋ encoder ደካማ መሆኑ እንደሚችል ይጠቁማሉ።

ጥቅሞቹ እንዳይታዩ — ወይም በስህተት እንዲታዩ — በቂ ትንሽ ነበሩ። ብዙ ውጤቶች የታዩት ከተለመደው በላይ ትልቅ sample እና ጥንቃቄ ያለው ስታቲስቲክ ስለተጠቀሙ ነው። ደራሲዎቹ በግልጽ ቃል፣ በውጤት መጠኑና በnoise ምክንያት ብዙ robotics ጽሑፎች የስታቲስቲክ ጫጫታን እየለኩ ሊሆን የሚችል ከፍተኛ አደጋ አለ ይላሉ። እንዲሁም ቀላል የሚመስል የዳታ *normalisation* ምርጫ ከarchitecture ለውጦች የበለጠ ውጤት እንዳሳደረ አገኙ፣ እና በpretraining ውስጥ የnormalisation bug የታወቀው evaluations ከተጠናቀቁ *በኋላ* ነበር።

ይህ ምናልባት ምን ማለት ነው?

የሚደገፈው ትርጉም፤ በተለያዩ የሮቦት ዳታዎች ላይ ትልቅ መጠን pretraining ማድረግ እውነተኛና ጠቃሚ ንጥረ ነገር ነው። ለአዲስ ሥራ ያነሰ ዳታ ያስፈልጋል፣ እና እውነተኛው ዓለም ከስልጠና ጋር ሳይመሳሰል policyውን ይበልጥ ጽኑ ያደርገዋል። ይህ መስኩ እየተጫወተበት ያለውን አቅጣጫ በእውነት ይደግፋል። ጥቅሞቹ ግን *መጠነኛ* እና *በሁኔታ* የተገደቡ ናቸው፤ በአብዛኛው ከfinetuning በኋላ ይታያሉ፣ በሥራዎች ሁሉ በአንድ ላይ ሲመዘኑና በስርጭት shift ሲፈተኑ ይበልጥ ግልጽ ናቸው። ይህ በቀጥታ ለማንኛውም ሥራ የሚሄድ አጠቃላይ ሮቦት መምጣት አይደለም።

የበለጠ ጸጥ ያለው ግን ጠቃሚ ትርጉም የmethodology ነው። ጥናቱ እንደ መለኪያ መስፈርት ይሠራል፤ ስለ robot policy ሊታመን የሚችል አባባል ለማድረግ ምን ያህል ማስረጃ እንደሚያስፈልግ ያሳያል፣ እና ብዙ የታተሙ አስደናቂ ውጤቶች በቂ ማስረጃ ላይ ላይመሠረቱ እንደሚችሉ ይጠቁማል።

ይህ ጥናት የማያረጋግጠው

- **አጠቃላይ-ዓላማ ሮቦት አይደለም።** ጥቅሞቹ በተወሰነ architecture — ዲፊውዥን policies — ላይ፣ ለእያንዳንዱ ሥራ finetune ተደርጎ፣ በተቆጣጠሩ አካባቢዎች እና ከteleoperated demonstrations ተገኝተዋል። በትዕዛዝ ማንኛውንም አዲስ ሥራ የሚሠራ autonomous robot አይደለም።
- **Zero-shot አጠቃቀምን አያረጋግጥም።** Finetuning ሳይኖር ትልቁ ሞዴል single-ተግባር baselinesን በተደጋጋሚ አላሸነፈም።
- **"Emergent leap" ማስረጃ አይደለም።** ዳታ ሲጨምር አፈጻጸሙ በቀጣይነት ተሻሻለ፣ ድንገተኛ ዝላይ አልታየም።
- **ቁጥሮቹ relative እና lab-bound ናቸው።** Absolute success መጠኖች ንጽጽሩ ተጋላጭ እንዲሆን በፍላጎት ~50% አካባቢ ተዘጋጅተዋል፤ ይህ የበእውነተኛ ዓለም reliability መለኪያ አይደለም።
- **ማንኛውም policy ለምን እንደሚሳካ ወይም እንደሚወድቅ አይወስንም፤** LBM የባሰ ውጤት ያሳየባቸው አንዳንድ ሥራዎች ተመዝግበዋል ግን ምክንያታቸው አልተገለጸም።
- **ከግምገማ rig ውጭ ስለ ደህንነት፣ autonomy ወይም deployment ምንም አይልም።**

ማስረጃው ምን ያህል ጠንካራ ነው?

ዋና የንጽጽር አባባሎቹ — finetuned LBMs በአጠቃላይ from-scratch baselinesን እንደሚበልጡ፣ ብዙ ጊዜ ያነሰ ዳታ እንደሚፈልጉ፣ እና በስርጭት shift ስር ይበልጥ ጽኑ እንደሆኑ — ጠንካራና ከተለመደው በላይ በጥሩ ሁኔታ ተቆጣጥረው ተፈትነዋል። Blind፣ randomised፣ large-sample፣ statistically tested፣ እና በscoring ላይ QA አለ። ዋና ውጤቶቹን በቅርብ ወደ ቀጥተኛ ትርጉማቸው ለመቀበል በቂ የmethodology ጥራት አለ።

ደራሲዎቹ ራሳቸው የሚጠቅሷቸው ገደቦች ግን አሉ። ስህተት bars የግምገማ የዘፈቀደነትን ይይዛሉ፣ የስልጠና የዘፈቀደነትን ግን **አያካትቱም**። ተመሳሳይ ሞዴሉን ሁለት ጊዜ ካሰለጠኑ በትርጉም የሚለያይ policy ሊያገኙ ይችላሉ፣ ይህ ልዩነት በስታቲስቲኩ ውስጥ የለም። በእውነተኛ world 50 ሙከራዎች በእያንዳንዱ ሥራ መካከለኛ ውጤቶችን ለማግኘት በቂ ናቸው፣ ግን ትንሽ ውጤቶችን ሊያመልጡ ይችላሉ። ቋንቋ conditioning ትንሽ encoder ተጠቅሟል፣ ስለዚህ ከትልቅ ስርዓቶች ጋር "በቃ ለሮቦቱ ንገረው" የሚለው ጥያቄ ሊለይ ይችላል። ከግምገማ በኋላ የተገኘው የnormalisation bugም በግልጽ ተዘግቧል። እነዚህ ዋና ውጤቶቹን አያፈርሱም፣ ግን ጥናቱ መስኩ ብዙ ጊዜ ችላ የሚላቸው የዚህ ዓይነት ዝርዝሮች እንደሆኑ ይከራከራል።

አንድ የምንጭ ማስታወሻ፤ ይህ ማብራሪያ በደራሲዎቹ preprint ላይ የተመሠረተ ነው። በjournal የታተመውን እትም ማግኘት አልቻልንም፣ ስለዚህ በpreprint እና published text መካከል ለውጦች ካሉ አልመረመርንም።

ለምን ይጠቅማል?

"Robot foundation ሞዴሎች" ለማጋነን ቀላል የሆነ ሀረግ ነው። ይህን ጥናት "ሁሉም ነገር ሠራ!" ብሎ ማንበብም፣ "በሙሉ overhyped ነው" ብሎ መጥላትም ቀላል ነው። ትክክለኛው ትርጉም ከሁለቱም የበለጠ ጠቃሚ ነው፤ በተለያየ ዳታ pretraining እውነተኛ፣ ሊለካ የሚችል ግን መጠነኛ ጥቅም ይሰጣል — በተለይ ለእያንዳንዱ ሥራ ያነሰ ዳታ እና ይበልጥ ጽናት — እና ዳታ ሲጨምር መንገዱ በሚተነበይ ሁኔታ ይሻሻላል።

ጥናቱ የበለጠ አስፈላጊ የሚሆነው ጥንቃቄውን በራሱ መስክ ላይ ስለሚያደርግ ነው። እውነተኛ ውጤቶች በደካማ ግምገማ ውስጥ እስኪጠፉ ድረስ ትንሽ መሆናቸውን፣ እና እንደ መረጃ normalisation ያለ የተለመደ ምርጫ ከብልህ architecture ለውጥ የበለጠ ተጽእኖ ሊኖረው እንደሚችል ያሳያል። ስለዚህ ብዙ የrobot-learning "እድገት" ከመታመኑ በፊት ይበልጥ ጠንካራ መለኪያ ያስፈልገዋል።

ንጹህ ማጠቃለያ

Toyota Research Institute ተመራማሪዎች **large ባህሪ ሞዴሎች** — በ~1,700 ሰዓት የተለያዩ የሮቦት manipulation ዳታ ላይ ቀድሞ የተሰለጠኑ ዲፊውዥን-based policies — ገንብተው ከfrom-scratch single-ተግባር policies ጋር በከፍተኛ ጥንቃቄ አነጻጸሩ። ግምገማው blind፣ randomised እና large-sample ነበር — ≈1,800 በእውነተኛ ዓለም እና 47,000+ ማስመሰል ሙከራዎች — እና ትክክለኛ ስታቲስቲክ ተጠቅሟል። ለእያንዳንዱ ሥራ finetune ከተደረጉ በኋላ ትልቁ ሞዴሎች በአጠቃላይ የተሻለ ሠሩ፣ **3–5× ያነሰ ተግባር-specific መረጃ** ፈለጉ፣ እና **ሁኔታ ሲለወጥ ይበልጥ ጽኑ** ነበሩ። አፈጻጸም ከpretraining መረጃ ጋር በቀጣይነት ጨመረ። ነገር ግን **finetuning ሳይኖር** በተደጋጋሚ single-ተግባር ሞዴሎችን አላሸነፉም፤ ብዙ ውጤቶች ትንሽ ነበሩ፣ እና መረጃ normalisation ከarchitecture ይልቅ ትልቅ ተጽእኖ አሳደረ። ይህ robot-foundation-ሞዴል አቅጣጫን በመጠነኛ እና በሚለካ መልኩ የሚደግፍ ማስረጃ ነው — **አጠቃላይ ሮቦት አይደለም፣ zero-shot generalist አይደለም፣ "emergent leap"ም አይደለም** — እና

ብዙ robotics ጥናቶች ጫጫታ ሊለኩ እንደሚችሉ ጠንካራ ማስጠንቀቂያ ነው።

ያለ ማጋነን ማረጋገጫ

ጥናቱ የሚያሳየው: በ≈1,800 በእውነተኛ ዓለም + 47,000+ ማስመሰል rollouts የተደረገ blind፣ statistically-powered ግምገማ ላይ፣ multitask-pretrained-then-finetuned ዲፊውዥን policies (LBMs) በአጠቃላይ from-scratch single-ተግባር policiesን ይበልጣሉ፣ ~3–5× ያነሰ ተግባር-specific መረጃ በመጠቀም ተመሳሳይ አፈጻጸም ይደርሳሉ፣ እና በስርጭት shift ይበልጥ ጽኑ ናቸው። አፈጻጸሙ ከpretraining መረጃ ጋር በቀጣይነት ይጨምራል።

ምክንያታዊ ግን ያልተረጋገጠ: እነዚህ ጥቅሞች ወደ በጣም ትልቅ vision-ቋንቋ-action ሞዴሎች መሸጋገራቸው፤ እና የታየው ለስላሳ scaling ከተፈተነው ዳታ ክልል በላይ መቀጠሉ።

የማያሳየው: አጠቃላይ-ዓላማ ወይም zero-shot robot፤ ድንገተኛ "emergent" የችሎታ ዝላይ፤ የተወሰኑ ሥራዎች ለምን እንደሚወድቁ ማብራሪያ፤ በabsolute terms የበእውነተኛ ዓለም reliability፤ ስለ ደህንነት ወይም autonomous deployment ማስረጃ።

ዋና ገደቦች: ስታቲስቲኩ የግምገማ የዘፈቀደነትን ይይዛል እንጂ የስልጠና-run የዘፈቀደነትን አይይዝም፤ 50 በእውነተኛ ዓለም ሙከራዎች ትንሽ ውጤቶችን ሊያመልጡ ይችላሉ፤ አንድ architecture እና አንድ lab፤ መጠነኛ ቋንቋ encoder፤ ከevaluations በኋላ የተገኘ መረጃ-normalisation bug፤ ትንታኔው በpreprint ላይ የተመሠረተ ነው።

አጠቃላይ አንባቢ ምን ያህል እምነት ሊኖረው ይገባል? Multitask pretraining + finetuning እውነተኛ ግን መጠነኛ ጥቅሞች — በተለይ የዳታ ውጤታማነትና ጽናት — እንደሚሰጥ ከፍተኛ እምነት፣ እና እነዚህ በከፍተኛ ጥንቃቄ እንደተለኩ ከፍተኛ እምነት። ይህ **አጠቃላይ ወይም** zero-shot robot **አለመሆኑ** እና emergent leap **አለመሆኑ** ላይም ከፍተኛ እምነት። ወደ ትልቅ ሞዴሎች ምን ያህል እንደሚስፋፋ ላይ መካከለኛ እምነት። የደራሲዎቹ ዋና ማስጠንቀቂያ ግን በጥብቅ ሊወሰድ ይገባል፤ ውጤቶቹ ትንሽ ስለሆኑ ደካማ ኃይል ያላቸው ጥናቶች የስታቲስቲክ ጫጫታን እንደ እድገት ሊያቀርቡ ይችላሉ።

ምንጮች

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/