The Clean Paper

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

Büyük robot "foundation modelleri" gerçekten daha mı iyi çalışıyor? Dikkatli cevap: ölçülü biçimde evet — ve çoğu çalışma bunu ayırt edemeyebilir

Toyota Research Institute, yaklaşık 1.700 saatlik çeşitli manipülasyon verisiyle ön-eğitilmiş "large behavior model" robot politikalarını sıfırdan tek-görev politikalarına karşı alışılmadık titizlikle test etti: kör, randomize, büyük örneklemli (~1.800 gerçek dünya, 47.000+ simülasyon) ve gerçek istatistik kullanan denemeler. Görev başına fine-tuning sonrasında büyük modeller ortalamada daha iyi performans gösterdi, yaklaşık 3–5× daha az göreve özgü veri gerektirdi ve koşullar değiştiğinde daha dayanıklıydı; daha fazla pretraining verisiyle performans düzgün biçimde yükseldi. Fakat fine-tuning olmadan tek-görev modellerini tutarlı biçimde geçmediler, bazı etkiler ancak büyük örneklemle görülecek kadar küçüktü ve sıradan bir veri-normalizasyon tercihi mimariden daha fazla önem taşıdı. Robot-foundation-model yönü için ölçülü destek — genel amaçlı robot değil, zero-shot genelci değil, "emergent sıçrama" değil — ve robotik çalışmalarının önemli bölümünün gürültü ölçüyor olabileceğine dair açık bir uyarı.

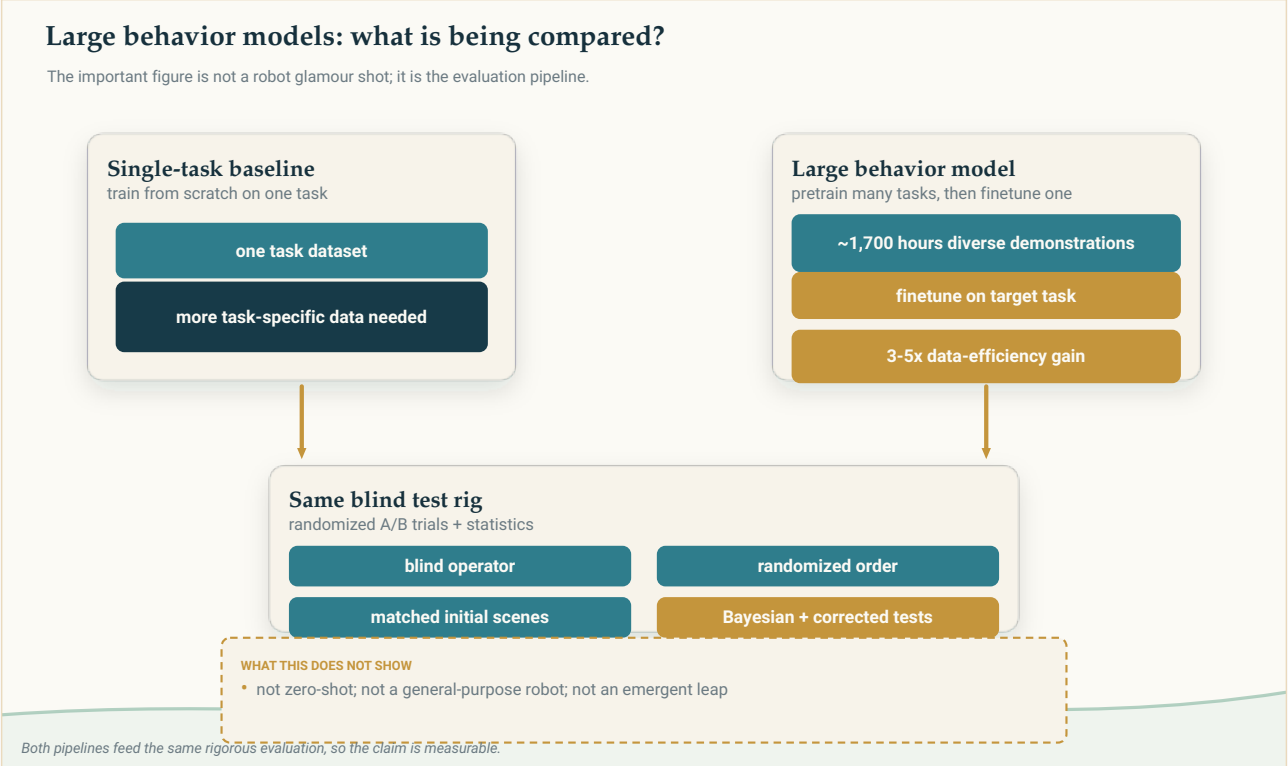
Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

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

Asıl konusu dürüstlük olan bir robot makalesi

Robotik şu anda bir "foundation model" yarışının ortasında. Dil ve görüntü yapay zekâsından ödünç alınan fikir cezbedici: robotu her görev için ayrı ayrı eğitmek yerine, tek bir büyük **"large behavior model" (LBM)** modelini devasa ve çeşitli bir gösterim yığını üzerinde eğit ve çok farklı işlere uyum sağlayabilen, hızlıca özelleşen bir sistem elde et. Heyecan — ve yatırım — muazzam. Başlık kendi kendini yazıyor: *genel amaçlı robot beyinleri geldi.*

Toyota Research Institute'tan gelen bu makale tam da o başlığı yazmayı reddettiği için ilginç. Asıl katkısı daha gösterişli bir robot değil. Aldatıcı biçimde basit bir soruya sıkı biçimde bakıyor: *büyük-model yaklaşımı gerçekten daha mı iyi çalışıyor ve bunu nasıl anlayabiliriz?* Yazarların kendi ifadesiyle alan için alışılmadık düzeyde istatistiksel özenle verilen cevap, gerçekten yararlı bir "evet, ama" ve robotik çalışmalarının önemli bir bölümünün gürültü ölçüyor olabileceğine dair bir uyarı.



Her iki yaklaşım aynı kör, randomize değerlendirmeye giriyor. Makalenin iddiası robot foundation modellerinin zero-shot genelci olduğu değil; pretraining'in dikkatli ölçüldüğünde veri verimliliğini ve dayanıklılığı artırabileceği.

Yazarlar ne yaptı?

LBM'leri somut ve belirli bir anlamda kurdular: **diffusion tabanlı visuomotor politikalar** (kamera görüntülerini, kısa bir dil talimatını ve robotun kendi eklem konumlarını okuyan, 10 Hz'de kısa

motor-komut dizileri üreten bir Diffusion Transformer). Bunlar **yaklaşık 1.700 saatlik** robot gösterimiyle — kurum içinde toplanmış 500'den fazla farklı görev ve kamuya açık veri setleri — **ön-eğitildi**, ardından tek tek görevlerde **fine-tuning** yapıldı. Çalışma boyunca karşılaştırma, aynı görevin verisiyle sıfırdan eğitilmiş **tek-görev politikasına** karşı.

Fakat makalenin kalbi *değerlendirme*. Yazarlar bunu asıl sonuç gibi ele alıyor. Kendilerini kandırmamak için şunları kullandılar:

- Gerçek dünyada **kör, randomize A/B testi** — robotu çalıştıran insan hangi politikanın test edildiğini bilmiyordu ve sıra rastgeleydi.
- **Kontrollü, tekrarlanabilir başlangıç koşulları** — her denemeden önce operatörler sahneyi bir görüntü overlay'ine eşleştirdi.
- **Büyük deneme sayıları** — gerçek dünyada her görev, politika ve koşul için 50 rollout; simülasyonda görev başına 200. Toplamda yaklaşık **1.800 kör gerçek-dünya rollout'u ve 47.000'den fazla simülasyon rollout'u**.
- **Doğru istatistik** — üst üste gelen hata çubuklarına göz kararı bakmak yerine başarı olasılığı için Bayesçi tahminler ve çoklu-karşılaştırma düzeltmeli ikili hipotez testleri. İnsan tarafından puanlanan denemelerin dörtte birinde puanlama hatasını ölçmek için ayrıca kalite güvence turu yaptılar.

[video pending]

Kahvaltı masası hazırlayan modellerin yan yana karşılaştırması: solda tek-görev baseline, sağda LBM. İki video da 1× hızda oynuyor. Bu değerlendirilen tek bir görevdir; genel amaçlı özerkliğin kanıtı değildir.

Asıl mesele bu düzenek. Makalenin bütünü, böyle bir değerlendirme olmadan gerçek iyileşmeyle şansı birbirinden ayıramayacağınızı savunuyor.

Ne buldular?

Fine-tuning yapılmış büyük modeller, sıfırdan eğitilmiş tek-görev modellerini ortalamada geçti. Görevler birlikte ele alındığında, ön-eğitilmiş ve ardından belirli görevde fine-tuning yapılmış LBM, hem simülasyonda hem gerçek dünyada aynı görev için sıfırdan eğitilmiş politikadan güvenilir biçimde daha iyi performans gösterdi ve fark istatistiksel olarak anlamlıydı. Tek tek görevlerde de fine-tuned LBM neredeyse her durumda istatistiksel olarak en az sıfırdan-eğitim kadar iyiydi (gerçek dünyada 3/3 görev, simülasyonda 15/16).

En büyük ve en net kazanç veri verimliliği. Fine-tuned bir LBM, sıfırdan-eğitimle eşdeğer performansa yaklaşık **3–5× daha az göreve özgü veriyle** ulaştı. Gerçek dünyadaki bir görevde (kahvaltı masası hazırlama), gösterimlerin yalnızca **%15'i** üzerinde fine-tuning yapılan LBM, gösterimlerin **%100'üyle** sıfırdan eğitilen politikayı geçti.

Pretraining en çok koşullar değiştiğinde yardımcı oluyor. Test ortamı eğitim koşullarından kasıtlı biçimde uzaklaştırıldığında ("distribution shift"), fine-tuned LBM'nin avantajı *büyüdü*. Bir

simülasyon kümesinde normal koşullarda 16 görevin 3'ünde sıfırdan-eğitimi istatistiksel olarak geçerken, distribution shift altında 16 görevin 10'unda geçti. Gerçek konuşlandırmalar eğitim koşullarından daima bir miktar uzaklaştığı için bu dayanıklılık muhtemelen pratik açıdan en önemli bulgu.

Daha fazla pretraining verisi düzgün biçimde yardımcı oldu. Pretraining verisi arttıkça performans istikrarlı biçimde yükseldi — test edilen ölçeklerde ani bir sıçrama veya "emergent" atlama **yoktu**. Yararlı, öngörülebilir ve dramatik olmayan bir ölçekleme.

Fakat fine-tuning olmadan genelci model hikâyesi tutmadı. Ön-eğitilmiş LBM *zero-shot* — göreve özel fine-tuning olmadan — kullanıldığında tek-görev politikalarını tutarlı biçimde geçmedi. Tek bir ağ aynı anda birçok görev yapabiliyordu, fakat "sadece ne yapacağını söyle" hayali burada doğrulanmadı; yazarlar bunun bir bölümünü küçük dil encoder'larının kırılganlığına bağlıyor.

Ve kazançlar kaçırılması — hatta sahte görünmesi — kolay olacak kadar küçüktü. Etkilerin çoğu ancak alışılmıştan büyük örneklem sayıları ve dikkatli testlerle görünür hâle geldi. Yazarlar, etkilerin boyutu ve gürültü göz önüne alındığında **robotik makalelerinin önemli bir bölümünün istatistiksel gürültüyü ölçüyor olma riski bulunduğunu** açıkça söylüyor. Ayrıca sıradan bir tercihin — verinin nasıl *normalleştirildiğinin* — mimari değişikliklerden daha fazla sonuç etkilediğini ve pretraining'deki bir normalizasyon **hatasının** değerlendirmeler bittikten *sonra* ortaya çıktığını buldular.

Bu büyük olasılıkla ne anlama geliyor?

Savunulabilir yorum şu: çeşitli robot verileri üzerinde büyük ölçekli pretraining gerçekten işe yarayan ve değerli bir bileşen — yeni görev başına daha az veri gerektiriyor ve dünya eğitimle birebir eşleşmediğinde politikaları daha dayanıklı yapıyor. Bu, alanın yatırım yaptığı yönü gerçekten destekliyor. Fakat kazançlar *orta düzeyde ve koşullu* (çoğunlukla fine-tuning sonrasında ortaya çıkıyor, toplu sonuçlarda ve stres altında en net), hazır kullanımlı genel robotun gelişi değil.

Daha sessiz ve belki daha önemli anlam metodolojik. Makale bir anlamda ölçüm cetveli: robot politikası hakkında güvenilir iddia kurmak için gerçekte ne kadar kanıt gerektiğini gösteriyor ve yayımlanan heyecanın önemli bir kısmının fazla az veriye dayanabileceğini ima ediyor. Alanın başka bir modelden daha çok ihtiyaç duyduğu düzeltme bu olabilir.

Bu neyi *kanıtlamıyor*?

- **Genel amaçlı bir robot değil.** Kazançlar belirli bir mimaride (diffusion politikaları), görev başına fine-tuning ile, kontrollü koşullarda ve teleoperasyon gösterimlerinden elde edildi — komutla keyfî yeni işler yapan özerk robot değil.
- **Zero-shot** kullanımı doğrulamıyor. Fine-tuning olmadan büyük model tek-görev baseline'larını tutarlı biçimde geçmedi.
- **"Emergent sıçrama"** kanıtı değil. Ölçekleme performansı düzgün biçimde artırdı; "bir noktada

birdenbire yetenek kazandı" anlatısını destekleyen kesinti yok.
- Sayılar **göreli ve laboratuvara bağlı**. Mutlak başarı oranları karşılaştırmaları hassaslaştırmak için bilinçli olarak yaklaşık %50 civarında ayarlandı; bunlar gerçek dünya güvenilirliğinin ölçümü değil ve çalışma tek laboratuvardan tek mimari.
- Herhangi bir politikanın **neden** başarılı veya başarısız olduğunu açıklamıyor; büyük modelin *daha kötü* olduğu birkaç görev raporlanıyor ama açıklanmıyor.
- Değerlendirme düzeneği dışında **güvenlik, özerklik veya konuşlandırma** hakkında hiçbir şey söylemiyor.

Kanıt ne kadar güçlü?

Temel karşılaştırmalı iddialar için — fine-tuned LBM'ler toplu olarak sıfırdan baseline'ları geçiyor, birkaç kat daha az veri gerekiyor ve distribution shift altında daha dayanıklılar — kanıt hem güçlü hem alışılmadık derecede iyi kontrollü: kör, randomize, büyük örneklemli, istatistiksel testli ve puanlamada QA kontrollü. Metodolojinin başlıktaki sonuçları neredeyse doğrudan kabul etmeye yetecek kadar sağlam olduğu nadir durumlardan biri.

Dürüst çekinceler, yazarların kendilerinin öne sürdüğü şeyler. Hata çubukları *değerlendirme* rastlantısallığını kapsıyor ama **eğitim rastlantısallığını kapsamıyor** — aynı modeli iki kez eğitmek anlamlı biçimde farklı politika üretebilir ve bu varyasyon istatistikte yok. Gerçek dünya görevlerinde 50'şer deneme orta büyüklükte etkileri yakalamaya yeterli, küçük etkileri kaçırabilir. Dil-koşullandırma mütevazı bir encoder kullandı; "robota sadece ne yapacağını söyle" iddiaları daha büyük sistemlerde farklı olabilir. Bir de değerlendirme sonrasında bulunan normalizasyon hatasının açıkça bildirilmesi var. Bunların hiçbiri ana bulguları batırmıyor; tam tersine makalenin alanın genellikle görmezden geldiğini söylediği türden ayrıntılar bunlar.

Aynı ruhla bir kaynak notu: bu açıklama yazarların preprint'ine dayanıyor. Dergide yayımlanan sürüme erişemedik, bu nedenle preprint ile yayımlanmış metin arasında değişiklik olup olmadığını kontrol etmedik.

Neden önemli?

"Robot foundation modelleri" aşırı iddia için üretilmiş bir ifade gibi ve böyle bir çalışma iki yönde de kolay yanlış okunabilir — zafer ilan eden *"çalışıyor!"* ya da küçümseyen *"abartılıyor."* Doğru yorum ikisinden de daha kullanışlı: çeşitli veriler üzerinde pretraining gerçek, ölçülebilir ama orta büyüklükte yararlar sağlıyor — başlıca görev başına daha az veri ve daha fazla dayanıklılık — ve yol ölçekle öngörülebilir biçimde iyileşiyor.

Daha derin önem, makalenin bu titizliği kendi alanına çevirmesinde. Gerçek etkilerin özensiz değerlendirmede kaybolacak kadar küçük olduğunu ve veri normalizasyonu gibi sıkıcı bir seçimin akıllı yeni mimariden daha büyük fark yaratabileceğini göstererek robot öğrenme ilerlemesinin önemli bölümünün inanılmadan önce daha sağlam ölçüm gerektirdiğini savunuyor. Güvenilirliğini "sonuç"

ile “dilek” arasındaki farkı korumak için harcayan bir makale, liderlik tablosunda birinci olmaktan daha nadir ve daha değerli bir şey yapıyor.

Kısa özet

Toyota Research Institute araştırmacıları “large behavior model” denen, yaklaşık 1.700 saatlik çeşitli manipülasyon verisiyle ön-eğitilmiş diffusion tabanlı robot politikaları geliştirdi ve bunları sıfırdan eğitilmiş tek-görev politikalarına karşı alışılmadık derecede titiz bir protokolle test etti: kör, randomize, büyük örneklemli (≈1.800 gerçek-dünya ve 47.000+ simülasyon denemesi) ve gerçek istatistik kullanan değerlendirme. Görev başına fine-tuning sonrasında büyük modeller toplu olarak güvenilir biçimde daha iyi performans gösterdi, yaklaşık **3–5× daha az göreve özgü veri** gerektirdi ve **koşullar değiştiğinde daha dayanıklıydı**; pretraining verisi arttıkça performans düzgün biçimde yükseldi. Ancak **fine-tuning olmadan** tek-görev modellerini tutarlı biçimde geçmediler, bazı etkiler ancak büyük örneklem sayesinde görülecek kadar küçüktü ve sıradan bir veri-normalizasyon tercihi mimariden daha fazla önem taşıdı. Robot-foundation-model yönü için sağlam ve ölçülü destek — **genel amaçlı robot değil**, zero-shot genelci değil ve “emergent sıçrama” değil — ayrıca robotik çalışmalarının önemli kısmının gürültü ölçüyor olabileceğine yönelik güçlü bir uyarı.

Abartısız değerlendirme

Makalenin gösterdiği: Titiz, kör ve istatistiksel gücü yüksek değerlendirmeyle (≈1.800 gerçek dünya + 47.000+ simülasyon rollout’u), çok-görevli pretraining ardından fine-tuning yapılan diffusion politikaları (LBM’ler) toplu olarak sıfırdan tek-görev politikalarını geçiyor, yaklaşık 3–5× daha az göreve özgü veriyle eşdeğer performansa ulaşıyor ve distribution shift altında daha dayanıklı; performans pretraining verisiyle düzgün biçimde ölçekleniyor.

Makul ama kanıtlanmamış olan: Bu yararların çok daha büyük vision-language-action modellere taşınacağı (dil encoder’ları küçüktü); düzgün ölçeklemenin test edilen veri aralığının ötesinde sürdüğü.

Göstermediği: Genel amaçlı veya zero-shot robot (fine-tuning yok → tutarlı avantaj yok); herhangi bir “emergent” yetenek sıçraması; belirli görev başarısızlıklarının açıklaması; mutlak gerçek-dünya güvenilirliği (başarı oranları duyarlılık için ~%50’ye ayarlandı); güvenlik veya özerk konuşlandırmayla ilgili herhangi bir şey.

Başlıca sınırlamalar: İstatistik değerlendirme rastlantısallığını kapsıyor, eğitim koşusu rastlantısallığını değil; görev başına 50 gerçek-dünya denemesi küçük etkileri kaçırabilir; tek mimari ve tek laboratuvar; mütevazı dil encoder’ı; değerlendirmeler sonrası veri-normalizasyon hatası bulundu; analiz preprint’e dayanıyor (yayımlanmış sürüm kontrol edilmedi).

Genel bir okur ne kadar güven duymalı? Çok-görevli pretraining + fine-tuning’in özellikle veri verimliliği ve dayanıklılıkta gerçek, orta büyüklükte faydalar verdiğine ve bunların alışılmadık dere-

cede dikkatli ölçüldüğüne yüksek güven. Bunun **genel amaçlı veya zero-shot robot olmadığına** ve **emergent sıçrama olmadığına** yüksek güven. Daha büyük modellere ne kadar ölçekleneceğine orta güven. Ve yazarların kendi uyarısını ciddiye almak gerekir: alanın etkileri, düşük istatistiksel güçlü çalışmaların yalnızca gürültü raporlamasına yetecek kadar küçük olabilir. Uygun tutum: yaklaşım konusunda ölçülü iyimserlik ve bu tür istatistiksel destek sunmayan robot-AI sonuçlarına sağlıklı şüphe.

Kaynaklar

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/