The Easy World

What the field's most used manipulation dataset actually contains, and why the way we measure physical AI has to change

Harry Khan, CEO, Physicore
10 August 2026

Summary

We reviewed every run in the public sample of DROID, the manipulation dataset behind flagship models including MolmoAct 2 and pi05, and assessed the dataset against 57 questions covering documentation, coverage, labelling, failure representation, benchmark fitness, and rights.

What we found:

- **58 of 100 runs carry no task instruction.** There is no stated objective, so no reviewer and no training pipeline can determine what the run was attempting or whether it succeeded.
- **No run carries an outcome marker, and 14 of 100 are outright failures.** A dropped glass, a knocked-over cup, failed grasps of a bottle, a remote and a yoghurt pot, an unsuccessful keypress, one recording of zero usable length. Nothing separates them from the successes beside them.
- **1 of 100 runs presents any real difficulty.** 47 clean, 47 normal, 91 brightly lit. Clutter, occlusion, poor light, damaged or reflective objects and human interference are effectively absent.
- **Recovery appears in 2 of 100 runs and is nowhere labelled.** The single behaviour a deployed robot most needs is both rare and invisible to any pipeline consuming the data.
- **Three sources report three different totals**, spanning roughly 17,600 episodes.
- **Published calibration corrections reach fewer than half of all episodes.**
- **The most downloaded copy declares a licence the makers did not grant.** The project homepage states no licence at all.

The field is measuring the easy world with confidence and the difficult world not at all. That is the finding, and it is why episode count has stopped being a useful measure of anything.

---

1. Why we looked here

Robot policies that perform well under evaluation degrade when they meet real conditions. The causes are known and they are all properties of data: narrow condition coverage,

demonstrations that contain successes but no corrections, and brittleness to variation that was never present during training.

Every one of those is a dataset problem, not a model problem. Which makes it remarkable that the datasets training the field's leading models have never been independently examined. They are documented by the teams that built them and checked by nobody.

We looked at DROID because it is the most consequential case available. Approximately 76,000 successful demonstrations, 564 scenes, 52 buildings, three continents, 50 operators, twelve months, identical hardware at thirteen institutions. It is a serious piece of work and it is in the training mix of the models the field is currently celebrating.

We were not looking for faults. We were establishing what the data actually contains, because nobody had.

---

2. How we assessed it

Two sources, held against each other.

What the makers state. The DROID project site and the accompanying paper.

What the data contains. All 100 runs of `lerobot/droid_100`, the sample in common circulation, reviewed individually via the LeRobot visualizer.

Each run was recorded on the task text shown, the observed action, the outcome, whether text and action agreed, camera condition, setting difficulty and lighting, and whether failure or recovery was visible. Setting difficulty was classed as Clean (single object, tidy surface, nothing obstructing), Normal (ordinary domestic or office conditions) or Challenging (clutter, occlusion, human interference or degraded visibility).

Alongside this, 57 questions were answered against the documentation, each with a source and a date. Where nothing existed, we recorded that nothing existed and where we searched.

---

3. What we found

3.1 Most runs have no objective

58 of 100 runs display no task text.

This is not a documentation gap. It is a hole in the middle of the data. Without a stated objective there is no way to determine what a run was attempting, which means there is no way to determine whether it succeeded.

It produced 21 runs we could not score at all. An arm lifts a towel and puts it down somewhere else. A container is picked up and returned to the same spot. Beads are moved onto paper, which is then folded. Completed task or abandoned attempt? The data does not say, and neither can anyone else.

A demonstration with no objective is not a demonstration. It is footage.

3.2 Failures sit unmarked among the successes

No run carries any outcome field. 14 of 100 are unambiguous failures on inspection.

A glass dropped while being moved into a dishwasher. A cup knocked over during an attempt to retrieve a pen. Failed grasps of a bottle, a television remote, a yoghurt pot. An unsuccessful attempt to press a key. One recording containing nothing usable at all.

None of these is distinguishable from a success in the data. Any pipeline treating this sample as demonstration material is teaching a model that dropping the glass is how the task is done.

Table 1. Outcomes on inspection (n = 100)

Outcome	Count
Success	62
Failed	14
Partial	3
Could not be determined	21

Table 2. Did the task text match the observed action? (n = 100)

Result	Count
Matched	42
Did not match	31
No text, or not assessable	27

3.3 Recovery is nearly absent, and invisible where it exists

Recovery appears in 2 of 100 runs. Neither is labelled.

In both, the arm failed an initial grasp and re-attempted successfully. That is exactly the behaviour a deployed robot needs when a grip slips on a wet surface or a package shifts. It is present twice in a hundred runs, and nothing in the data identifies it, so no training process can select for it or weight it even if it wanted to.

The field talks constantly about robustness. The data contains almost no examples of a robot recovering from its own mistake.

3.4 There is almost no difficulty in the data

1 of 100 runs presents challenging conditions.

Table 3. Setting difficulty (n = 100)

Classification	Count
Clean	47

Classification	Count
Normal	47
Challenging	1
Not recorded	5

Table 4. Lighting (n = 100)

Condition	Count
Bright	91
Dim	4
Not recorded	5

Objects are small graspable household and office items. The dominant structure is a single object moved from one point to another on a horizontal surface, in a kitchen or an office room, under even indoor light.

Clutter, occluded targets, reflective and damaged objects, degraded lighting, human presence, interruption: effectively unrepresented.

The conditions under which robots actually fail are missing from the data used to teach them.

3.5 A model can score well here without being any good

Given 94 of 100 runs in clean or normal settings and 91 brightly lit, with a repeating approach-grasp-move-release structure, a policy can post a strong number on this distribution by learning a narrow motion habit under near-constant visual conditions.

The score is real. What it proves is far smaller than it appears.

DROID is presented as an in-the-wild dataset. The scene count supports the breadth. Nothing in the sample supports the difficulty that phrase implies.

3.6 The totals do not agree

Table 5. Reported episode counts

Source	Total
DROID site and paper	~76,000 (successful only)
cadene/droid copy	92,223
MolmoAct 2 training report	74,604 usable after filtering

The 76,000 is success-only by design. Roughly 16,000 unsuccessful trajectories were released but excluded from the headline. The makers state this openly and the choice is defensible.

It is also why a dataset advertised as 76,000 demonstrations contains no failures in the part almost everyone uses.

3.7 Corrections reach a minority of the data

The makers described their original camera calibrations as noisy and published improved ones in April 2025. Those corrections cover approximately 36,000 of roughly 76,000 episodes.

Most of the dataset still carries the calibrations the makers themselves called noisy, and nobody downloading it is told.

3.8 The licence on the most downloaded copy is not the licence that was granted

The project homepage states no licence anywhere. The words do not appear on the page. The terms are traceable only through third-party publications, which record CC BY 4.0: commercial use permitted, attribution required.

The `lerobot/droid_100` copy states MIT, which is looser and requires no attribution.

Any company relying on that label believes it holds rights that were never granted to it. This is the most downloaded viewing copy of one of the most used datasets in robotics.

3.9 There is no consent or ethics statement to find

Human hands and bodies appear throughout, recorded in domestic and office interiors by around 50 operators, with language annotation crowdsourced to a commercial labelling platform.

We found no statement of consent, permission or release. No ethics or institutional review approval. No route for anyone appearing in the footage to ask for its removal.

We are not asserting that consent was not obtained. We are stating that no one downloading this data can establish that it was.

---

4. What DROID gets right

Three things, recorded because they are true.

The scene coverage is a genuine advance. 564 scenes across 52 buildings, against ten to twenty-four for the datasets that preceded it. This is the contribution, and nothing above reduces it.

The protocol discipline is rare. Identical arm, cameras and rig at all thirteen institutions. Community-collected data is very seldom this internally consistent.

The makers wrote down what they did. The collection method, the operator count, the success-only design, the later corrections: stated plainly. This assessment was possible because they were honest.

---

5. What this actually means

The individual findings matter less than the shape they make.

DROID measures short, single-object manipulation in tidy, well-lit rooms, and measures it competently. It does not measure readiness for deployment, because it contains almost none of the conditions that cause deployment to fail.

This is not a DROID problem. It is the shape of data collection across the field. Protocols optimise for volume, cleanliness and successful completion, because those are the things that are easy to collect, easy to count and easy to advertise. Difficulty is hard to produce. Failure is unwelcome. Correction is expensive to capture and nobody labels it.

So the field has built a measurement system that reliably certifies performance in favourable conditions, and then expresses surprise when models leave the room and stop working.

Two consequences follow directly.

Episode count has stopped meaning anything useful. Operators recorded up to a hundred trajectories per scene before moving on, so heavy near-repetition is designed into the corpus. Totals describe volume, not information.

The most important omission is not any single condition. It is that difficulty was never treated as something to capture at all.

---

6. What has to change

Performance on a dataset cannot be read as capability until the field agrees what data has to contain. On this evidence, the minimum is:

Every run has a stated objective. Runs without one are not demonstrations and should not be counted as such.

Every run has a graded outcome. Success, partial, failure, and the failure mode. A binary flag cannot tell a clean completion from a lucky one.

Failure and recovery are captured on purpose. The failed attempt and the correction after it, kept and labelled, not discarded and not left invisible.

Difficulty is a designed variable. Clutter, occlusion, variable and degraded lighting, reflective and damaged objects, human presence and interruption, objects placed away from the centre of the workspace. Present in known proportion, deliberately, not by accident.

Rights and provenance are stated where the data is obtained. Licence, ownership, consent basis and removal route, consistent across every copy.

Coverage is reported honestly. Distribution across scenes, tasks, conditions and operators published next to the headline number, so that information can be told apart from volume.

Physicore is building to this specification, and building the data that satisfies it: engineered stress testing in real operating environments, with the failure, the recovery and the correct demonstration all captured and labelled together.

This audit is why that is necessary.

---

7. Scope

Stated for precision, because every claim here should be checkable and exact.

The run-level findings describe all 100 runs of the lerobot/droid_100 sample, the copy in common circulation. They are findings about that sample and we make no proportional claim about the full corpus from them. The makers released expanded language annotations separately in December 2024; the finding at 3.1 is that the sample in general use does not carry them.

The documentation findings, at 3.6 through 3.9, concern the canonical release and its official materials.

Concentration, near-duplicate rate and train/evaluation overlap require analysis across all 76,000 episodes and are the subject of work now under way.

---

8. Verify it yourself

The full 57-question framework and the complete 100-run log are published alongside this report. Every finding can be checked against the DROID project site, the DROID paper and the lerobot/droid_100 sample as of August 2026.

If something here is wrong, tell us and we will correct it publicly.

References

Khazatsky, A. et al. *DROID: A Large-Scale In-the-Wild Robot Manipulation Dataset.* arXiv:2403.12945

DROID project site, droid-dataset.github.io

lerobot/droid_100, Hugging Face

cadene/droid, Hugging Face

MolmoAct 2 technical report, arXiv:2605.02881

Harry Khan, CEO, Physicore harry@physicore.ai · physicore.ai