ТЕХНИКАЛЫҚ ҒЫЛЫМДАР
ТЕХНИЧЕСКИЕ НАУКИ
TECHNICAL SCIENCE

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. No. 4 (32), Vol. 2. pp. 34–38
DOI: https://doi.org/10.5281/zenodo.23763950
Journal website: https://cajournal.kz

UDC 004.8

HOW MANY OPEN ROBOT DATASETS CAN ACTUALLY TRAIN A ROBOT? A CENSUS OF 400 LEROBOT DATASETS ON HUGGING FACE

Azamat Sagadat
School No. 215, Almaty, Kazakhstan

Azamat Sagadat — Grade 11 student, School No. 215, Almaty, Kazakhstan.

© Azamat Sagadat, 2026

Abstract: Robots learn manipulation from recorded demonstrations, and the LeRobot project has made it easy for anyone to record such demonstrations and share them on Hugging Face. On 23 October 2026, 81,881 datasets outside the LeRobot team's own account carried the LeRobot tag. This study asks how many of them can realistically be used to train a real robot. A uniform random sample of 400 was drawn and checked against a threshold proposed by the author, at least 10 recorded episodes, combined with three further filters: the name does not mark a test or an evaluation, there is at least one camera, and the name does not mark a simulation. Of the 389 datasets whose description could be read, 57% had ten episodes or fewer and the median number of episodes was 10. After these filters and a reading of the dataset cards, 148 datasets, 38% of those that could be read (95% confidence interval 33% to 43%), remained; for 9 of them a possible sign of simulation could not be resolved. The share depends on the threshold: it was 42% at 5 episodes and 30% at 20. Three of the 148 were re-uploads of large academic datasets that together held 95% of all episodes in the usable group; among the other 145, the median dataset had 48 episodes, 61% were recorded on the low-cost SO-100 or SO-101 arm family, and the ten largest held 68% of the episodes. In this sample, most uploads are too small, are tests or policy evaluations, or are simulations, and the usable part is dominated by one type of arm.

Keywords: robot learning, datasets, LeRobot, Hugging Face, data quality

1. Introduction

The author learned about LeRobot from a teacher who suggested that he explore the platform. He is interested in robotics and in how robots learn to perform real tasks. He checked random datasets, studied their episodes and judged whether they worked and could be used to train robots. He wanted to understand how useful openly available data really is for robotics.

Modern robot learning depends on data. Instead of programming every movement, researchers record many demonstrations of a task and train a model to reproduce them. Large shared collections have shown that pooling data helps: the Open X-Embodiment project assembled data from 22 different robots collected by 21 institutions, covering 527 skills, and found that a model trained on it improved the abilities of several robots [1]. Other large collections, such as DROID, with 76,000 demonstration trajectories from 564 scenes [4], and BridgeData V2, with 60,096 trajectories on a low-cost robot [5], were built by research groups with a common protocol. At the same time, the LeRobot library from Hugging Face [2] has made it possible for students and hobbyists to record demonstrations on cheap robot arms and upload them in a common format, and the number of public robot datasets has grown into the tens of thousands.

More data is only useful if it is good. A study of learning from human demonstrations found that results depend on the quality of the demonstrations [6], and researchers have argued that

datasets should come with documentation of what they contain and how they were made [3]. A dataset with one short recording, a test upload or a simulation that is not marked as one can mislead someone who wants to train a robot. This study asks a simple question: of the open robot datasets published with LeRobot, how many could be used to train a real robot, and what does the usable part look like?

2. The author's threshold and checks

The threshold used in this study was proposed by the author. After examining three small datasets from an earlier sample (9 October 2026), he wrote that a dataset with only two recordings is not enough, because the robot makes mistakes, and that "what matters is not the number of recordings but their quality and how well the robot shows itself in them. To be precise, I think 10 accurate recordings would be enough" (original in the Note). His threshold of 10 episodes became the first filter. It is a judgement rather than a standard from the literature, so the results are also reported for other thresholds.

The author then judged by hand, on an earlier pilot sample, five datasets that had passed the size, test and camera filters (12 October 2026): two were real robots suitable for training, two were simulations, and one could not be opened; its name identified it as a simulation as well. This showed that a simulation cannot be told apart from a real robot by its number of episodes, which led to the filter by name described below. He also judged four pilot datasets that recorded a reward signal (15 October 2026): two were real robots, one was a simulation and one could not be opened. A reward signal is typical for simulators but also appears in real robots trained by trial and error, so it was not used as a filter; datasets with a reward signal are reported as uncertain instead.

3. Data and methods

3.1. Population and sample

The list of all datasets tagged LeRobot was read from the Hugging Face Hub API on 23 October 2026 (endpoint /api/datasets with filter=LeRobot, all pages, 1,000 datasets per page). It contained 82,006 datasets; the 125 datasets of the LeRobot team's own account (lerobot/) were excluded, leaving 81,881 community datasets. A uniform random sample of 400 was drawn from the alphabetically sorted list with Python's random module (seed 7). For each dataset, the description file meta/info.json, which LeRobot creates automatically, was read. It gives the robot type, the number of episodes, the number of frames, the frame rate and the recorded streams. For 11 of the 400 datasets the file could not be found, leaving 389. The 400 datasets came from 332 different uploaders.

An earlier pilot sample, drawn on 5 October 2026, was used to develop the filters and for the author's hand checks. Its listing stopped after the first 10,000 datasets returned by the API, so it was not a sample of the whole population. On that pilot, about 32% of the readable datasets passed the same name, size and camera filters, fewer than in the new sample, most likely because the capped listing did not represent the full list.

3.2. Filters

The filters were applied in order: (1) at least 10 episodes; (2) the dataset name does not contain the strings test, eval, debug, tmp, temp, try, trial or demo; (3) at least one image or video stream; (4) the name does not contain a word that marks a simulator or a simulated benchmark (sim, simulation, mujoco, isaac, gym, metaworld, libero, pusht, robosuite, maniskill, genesis, pybullet, gazebo, robocasa, virtual, synthetic, minecraft, vpt, unity, sapien, habitat, calvin, mimicgen or robomimic). A camera is needed for policies that act from images; a dataset without one can still train a policy from joint positions alone. Among the datasets that passed all filters, those that recorded a reward or done signal or named no robot type were marked as uncertain, because these are the cases where a hidden simulation is most likely, and their dataset cards (README files)

© Central Asian Scientific Journal, 2026

were read. A dataset was removed when its own card described the data as simulated, scripted or rendered. The 95% confidence interval for the usable share was computed with the Wilson method.

4. Results

4.1. Most datasets are small

Of the 389 readable datasets, 120 (31%) had one episode or none, 100 (26%) had between 2 and 10, 65 (17%) had between 11 and 49 and 104 (27%) had 50 or more. In total, 220 (57%) had ten episodes or fewer. The median number of episodes was 10, and 33 datasets had exactly 10, so the threshold sits at a point where small changes matter. Of the readable datasets, 140 (36%) had a name containing one of the test words.

4.2. From 400 to 148

Table 1 shows how the filters reduced the sample. Of the 389 readable datasets, 187 had fewer than 10 episodes, 34 of the remainder had a test or evaluation name, 2 had no camera and 15 had a simulator in their name, such as libero, robocasa, isaaclab or Genesis. After these filters 151 datasets remained.

Table 1. Filtering a random sample of 400 LeRobot datasets

Step	Removed	Left
Random sample		400
Description could not be read	11	389
Fewer than 10 episodes	187	202
Name marks a test or evaluation	34	168
No camera	2	166
Simulation by name	15	151
Not a real robot, from the dataset card	3	148

Twelve of the 151 recorded a reward or done signal or named no robot type, and their cards were read. Three were not demonstrations recorded on a real robot, by their own description: scripted demonstrations by a program that reads the exact object positions ("privileged object poses") instead of a camera (6,000 episodes), robot and object meshes rendered over a 3D scene, and a task from the ALOHA simulation, whose name alohasim the name filter missed because the word sim is joined to another word. These three were removed, leaving 148 datasets, 38% of the readable ones (95% confidence interval 33% to 43%). For the other nine the cards gave no answer; if all nine were simulations, the usable share would fall to 36%.

The name filter for tests is rough. Of the 34 datasets it removed, 18 had eval or rollout in their name; in LeRobot these are usually recordings of a trained policy being tested, not human demonstrations, so removing them is justified. The other 16 matched test, demo or trial, at least two of them only as part of another word (test_tube, newdemos), and several, with 20 to 240 episodes, may be real recordings that kept a default name. Without the test filter, 182 datasets would pass.

Table 2. Usable datasets for different episode thresholds (same other filters)

Threshold	Usable	Share of 389
5 episodes	164	42%
10 episodes	148	38%
11 episodes	128	33%
20 episodes	118	30%
50 episodes	74	19%

4.3. What the usable part looks like

Three of the 148 usable datasets were re-uploads of large datasets collected by research groups: conversions of the KUKA and Fractal (RT-1) parts of Open X-Embodiment [1] and of DROID [4]. Together they held 389,325 of the 411,884 episodes in the usable group (95%). These are not community recordings, so the rest of this section describes the other 145.

The 145 community datasets came from 133 different uploaders. Of the 148 usable datasets, 89 (60%) were recorded on the SO-100 or SO-101 family, low-cost arms that people assemble themselves, in one-arm or two-arm versions. The median community dataset had 48 episodes and was recorded at 30 frames per second. Together the 145 held 22,559 episodes, 61% of the 37,229 episodes in all readable datasets other than the three conversions. The data were concentrated: the ten largest of the 145 held 68% of their episodes.

5. Discussion

In this sample, open robot data is large in number but much smaller in substance. More than half of the uploaded datasets have ten episodes or fewer, many are tests or recordings of policy evaluations, and some are simulations. Between a fifth and two fifths of the datasets that can be read pass all filters, depending on where the threshold is set, and a few of those may still be simulations.

The author's checks shaped the method. They showed that the number of episodes alone cannot separate real data from simulations, which led to the filter by name, and that a reward signal, which seems like a clear sign of simulation, also appears in real-robot data. Both points matter for anyone who builds automatic tools to judge datasets, and the cases in this census that only the dataset card could settle show that names and metadata alone are not enough.

The re-uploads of academic datasets show a second point. In episode counts, a handful of converted research datasets outweigh everything the community has recorded in this sample. Anyone counting the size of open robot data should separate the two, because they come from very different processes.

The dominance of the SO-100 and SO-101 arms reflects how LeRobot is used: it has made robot learning accessible with inexpensive hardware, so most community data comes from these arms. This is useful for people who own the same arm, but it also means that community data covers few kinds of robots. The results support the call for better documentation of datasets [3]: a dataset page that stated clearly whether the data is real or simulated, whether it is a test, and how many usable episodes it contains would save users from checking each one by hand.

5.1. Limitations

The sample of 400 is a small part of 81,881 datasets, and the results describe the sample. The threshold of 10 episodes is a judgement, and the counts change with it (Table 2). Name-based filters can miss tests and simulations with ordinary names and can remove real datasets with unusual names. Nine uncertain datasets could not be resolved from their cards, and datasets with ordinary names and no reward signal were not checked one by one, so the number of hidden simulations among the 148 is unknown. The 11 datasets without a description file may differ from the rest. Finally, the number of episodes says nothing about how well each episode was recorded, which the author named as the most important quality.

6. Conclusion

In a uniform random sample of 400 community LeRobot datasets, 148, or 38% of those that could be read, had at least 10 episodes, were not named as tests or simulations and had a camera. The share ranged from 19% to 42% depending on the episode threshold. Three converted research datasets held almost all the episodes; among community recordings, one low-cost arm family dominated and a few large datasets held most of the data. Open robot data is growing, but the share that can train a real robot is much smaller than the number of uploads suggests.

Note: original quotation

Section 2: «нужно не количество записей, но их качество и как в записях раскрывается робот. А если говорить четко думаю 10 точных записей хватит»

© Central Asian Scientific Journal, 2026

Data availability

The list of sampled datasets, their metadata and the analysis code are available from the author or his mentoring team on request. All datasets are public on the Hugging Face Hub.

Author contributions and use of AI tools

The author chose the research direction from a shortlist prepared by his mentor, proposed the threshold of 10 episodes and the idea that the quality of recordings matters more than their number, judged by hand, on earlier pilot samples, three small datasets, five datasets that passed the filters and four datasets with a reward signal, found that simulations cannot be recognised by size alone, and wrote the first paragraph of the Introduction in Russian, which his mentor translated. The data were collected and analysed and the text was drafted by his mentor with the help of a generative AI assistant (Claude, Anthropic). No number, quotation or reference was produced by a language model without being checked: every statistic is computed by the analysis scripts from the public data described above, and each statement about a cited work was checked against its published text or abstract.

Conflicts of interest

The author declares no conflicts of interest. No funding was received for this study.

References:

[1] Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models // Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). 2024. P. 6892-6903. https://doi.org/10.1109/ICRA57147.2024.10611477

[2] Hugging Face. LeRobot, GitHub repository. https://github.com/huggingface/lerobot (accessed October 2026).

[3] Gebru T., Morgenstern J., Vecchione B., Vaughan J.W., Wallach H., Daumé III H., Crawford K. Datasheets for datasets // Communications of the ACM. 2021. Vol. 64, No. 12. P. 86-92. https://doi.org/10.1145/3458723

[4] Khazatsky A., Pertsch K., Nair S., et al. DROID: A large-scale in-the-wild robot manipulation dataset // arXiv preprint. 2024. arXiv:2403.12945. https://arxiv.org/abs/2403.12945

[5] Walke H., Black K., Lee A., et al. BridgeData V2: A dataset for robot learning at scale // arXiv preprint. 2023. arXiv:2308.12952. https://arxiv.org/abs/2308.12952

[6] Mandlekar A., Xu D., Wong J., et al. What matters in learning from offline human demonstrations for robot manipulation // Conference on Robot Learning (CoRL). 2021. arXiv:2108.03298. https://arxiv.org/abs/2108.03298

© Central Asian Scientific Journal, 2026