Paper review · Empiricist

Robot concentration in Open X-Embodiment

Open X-Embodiment Collaboration, 2023 (arXiv:2310.08864).

Pavel Bushuyeu · November 2026

How many robots does OXE really cover?

The question

OXE pools 60 datasets from 22 robots. RT-X, Octo, OpenVLA and π0 train on it for transfer across robots.

Whether pooling helps a robot may depend on how much of the data comes from each robot, not only on how many robots are listed.

Hypothesis

A few robots and datasets hold most of the trajectories.

The mixtures that later models trained on reweight the datasets, but they stay concentrated on a few robots.

Public metadata only, one script

Sheet	OXE dataset overview and metadata tabs: robot, episodes, control rate, action space, language, operator.
TFDS	dataset_info.json and features.json for 69 datasets on gs://gresearch/robotics.
Mixtures	Octo OXE_MAGIC_SOUP and OpenVLA oxe_magic_soup_plus at pinned commits, with released dataset statistics.
Papers	OXE, Octo, OpenVLA and π0 text; the script checks all 23 quotes it uses.
Measures	Top-k share; effective number = 1 / sum of squared shares; sampling ∝ weight × steps, as both codebases do.

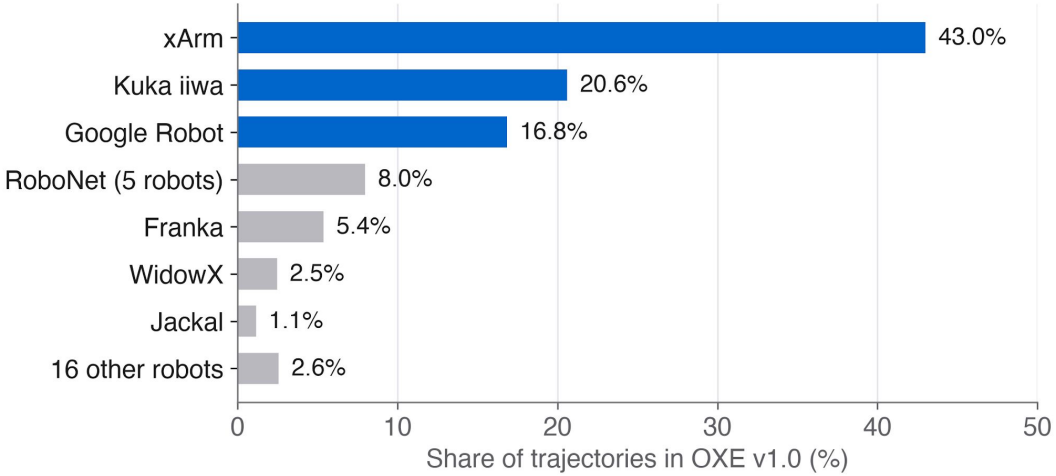
2.3 MB of metadata, no episodes. audit.py rebuilds every number and chart from the raw files.

The headline counts hold, with one fix

Claim	Paper	From the metadata
Datasets	60	60 rows marked v1.0
Robots	22	22 named; the figure likely counts RoboNet per robot and leaves out ALOHA
Trajectories	1M+	1.40M as listed; 1.03M with non-empty Kuka episodes
Labs, skills, tasks	34, 527, 160,266	Not in the metadata

The sheet lists QT-Opt (Kuka) with 580,392 episodes; the Octo and OpenVLA loaders count 209,880 non-empty ones. Figure 2(c) matches only the smaller number: xArm 43.0%, Kuka 20.6%, Google Robot 16.8%.

Three robots hold 80% of the trajectories



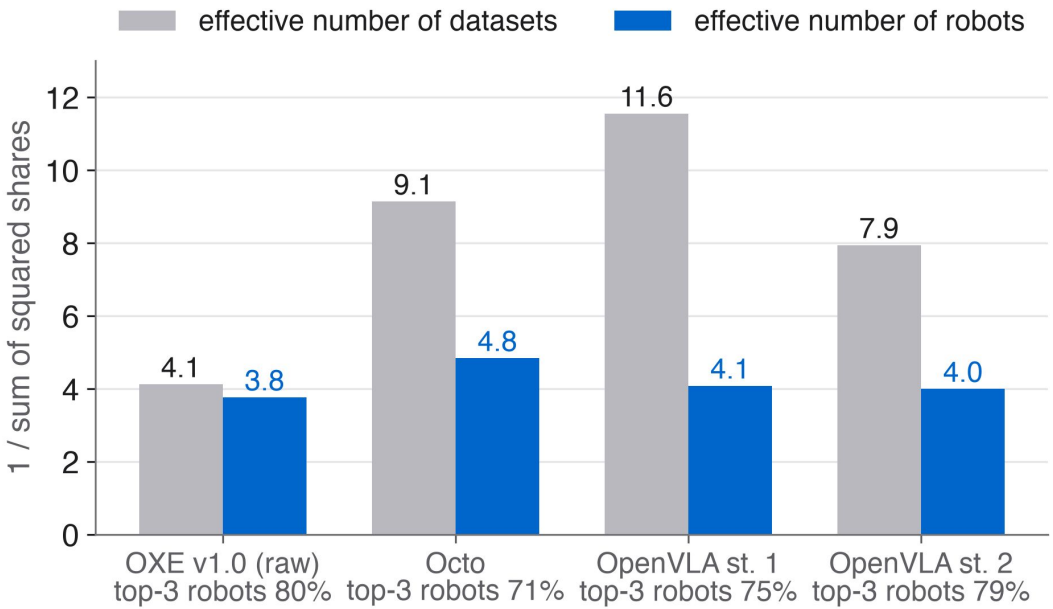
Effective number of robots: 3.8 of 23.

Language Table alone holds 43% of the trajectories.

Franka: 23 of 60 datasets, but 5.4% of the trajectories.

OXE v1.0, Kuka counted as non-empty episodes (as in Figure 2(c)).

Mixtures spread datasets, not robots



Reweighting doubles or triples the effective number of datasets.

The effective number of robots stays at 4.0 to 4.8.

Only 11 of the 22 robots are ever sampled.

Octo's table holds, OpenVLA's does not

Octo

Code weights × released dataset sizes give Octo's mixture table to within 0.03 points on average.

OpenVLA

Table 3 sums to 95%, which implies DROID at about 15% of the samples, not the stated 10%.

In the last third of training, the code drops Language Table and raises Fractal's weight from 0.54 to 1.0. One sample in four is RT-1 data.

Slow, templated and partly machine-driven

Share	OXE v1.0	Octo	OpenVLA st. 2
Control at 5 Hz or slower	15%	37%	44%
Control above 20 Hz	0.3%	2.4%	2.2%
No language label	50%	26%	22%
Templated language	12%	47%	59%
Human operator	47%	83%	86%

OXE v1.0: share of trajectories; Octo and OpenVLA: share of training samples. π0's own robots run at up to 50 Hz, and all its open-source data (OXE, Bridge v2, DROID) is 9.1% of its pre-training mixture.

What matches and what does not

Matches

60 datasets, 22 robots by count, 1M+ trajectories.

Franka has the most datasets and scenes.

xArm, Kuka and Google Robot give the most trajectories.

Octo's 27% of samples with a wrist camera.

Does not match

The 22 robots in Figure 2 are not the sheet's 22.

Figure 2(b): Google Robot has 4% of the listed scenes, not about 22%.

Octo's 56% of data with language.

π0 calls its OXE data both "the entire OXE dataset" and "OXE Magic Soup".

Diversity is in the datasets, not the robots

Challenges

Labels are contributors' form answers.

Step counts exist only for datasets in a mixture.

DROID's step count and π0's per-dataset weights are not published.

No episodes were opened, so the 370k missing QT-Opt episodes stay unexplained.

Insights

Cross-robot results from OXE rest on about four effective robots.

Reweighting moves samples between robots (xArm: 43% of trajectories, 6% of Octo's samples), but cannot add robots.

The open data is slow: half of it is at 10 Hz or less.

Robot share or dataset count?

Hypothesis: at a fixed step budget, a mixture balanced over robots transfers better to robots it under-samples than Octo's mixture does.

Train: two Octo-small models on the same steps, with Octo's weights and with an equal share per robot.

Test: SimplerEnv (Google Robot, WidowX) and LIBERO (Franka), plus few-shot fine-tuning on a held-out robot.

Measure: success with intervals, against each robot's share of the samples.

Sources

1. Open X-Embodiment Collaboration. Open X-Embodiment: Robotic Learning Datasets and RT-X Models, 2023.
2. Octo Model Team. Octo: An Open-Source Generalist Robot Policy, 2024; code github.com/octo-models/octo.
3. Kim et al. OpenVLA: An Open-Source Vision-Language-Action Model, 2024; code github.com/openvla/openvla.
4. Black et al. π0: A Vision-Language-Action Flow Model for General Robot Control, 2024.
5. OXE dataset overview sheet and TFDS metadata (gs://gresearch/robotics), fetched 2026-11-01.