What Would a GPT-3-Scale Robotics Dataset Cost?

Robotoken — A Preprint

Lucas Barbosa
NYU

29 September 2026

https://robotoken.lbxa.net

Abstract

We estimate the cost of collecting robotics data at language-model training scales through a five-year teleoperation program. The base acquisition cost is $219 per retained robot-hour. An illustrative 10 Hz accounting convention maps GPT-3's 300 billion training-token exposures to 8.3 million hours and about $1.8 billion; a two-trillion-token scenario costs about $12 billion, with a $3.1–44 billion range across cost assumptions. Public model disclosures provide scale comparisons, not evidence of equivalent robotics capability. Dedicated robot-data collection could require multibillion-dollar programs, but the evidence does not establish the necessary experience or prove that a single company cannot finance it.

Keywords Robot learning · Robotics data · Data scaling · Teleoperation

Reading the comparison. The selected 2T reference maps mechanically to 56 million hours; at base rates, collecting that additional private corpus costs about $12B and requires about 13,900 stations over five years. Illustrative $1B and $10B budgets buy about 4.6M and 46M retained hours. Applying the same convention to the 45T disclosure gives 1.25B hours and about $270B at base cost. These conversions do not establish a robotics capability target; public dataset durations are scale references, with different filtering, diversity and collection costs.

1 Define the milestone before counting data

The intended capability is broad manipulation: adapting to unfamiliar tasks, objects and environments from instructions or a few demonstrations. This excludes a separate estimate for humanoid locomotion. No source reviewed establishes how many hours deliver this capability, or a validated equivalence between robotics and GPT-3. GPT-3's actual training budget was **300 billion token exposures**, sampled from a corpus totaling about 499 billion tokens; some subsets were reused and others only partly sampled [1]. Two trillion is therefore a stress assumption, 6.67 times GPT-3's exposure count, not its historical training requirement.

Economic primitive. Use one retained hour of newly collected, nonduplicated robot experience: synchronized observations, robot state, actions and outcomes, with task/environment metadata. Retention requires rights, synchronization and quality checks; it does not guarantee novel skills or statistical independence. Useful failures may be retained. Repeated views and training epochs do not constitute new physical experience.

Optional token bridge. Define one *accounting transition* as an observation–action–next-state record at $f = 10$ Hz. This is a convenient convention borrowed from a prior estimate [8], not an assertion that 100 ms of motion contains one language token's information. Do not multiply the acquired experience by camera count, image patches or joint dimensions. For a token-exposure target T, mean reuse e and residual real-data share α after other sources:

$$H = \frac{\alpha T}{3{,}600\, f\, e}. \qquad \begin{array}{ll} T = 300 \times 10^9 & : H = 8.333 \text{ million hours}, \\ T = 2 \times 10^{12} & : H = 55.556 \text{ million hours}, \end{array} \tag{1}$$

where the displayed cases assume $\alpha = e = 1$ and $f = 10$.

This bridge is not invariant to tokenization. FAST encodes a one-second shirt-folding action chunk in an average of 53 tokens versus 700 with per-dimension binning, at comparable reconstruction accuracy [9]. Literal token parity would

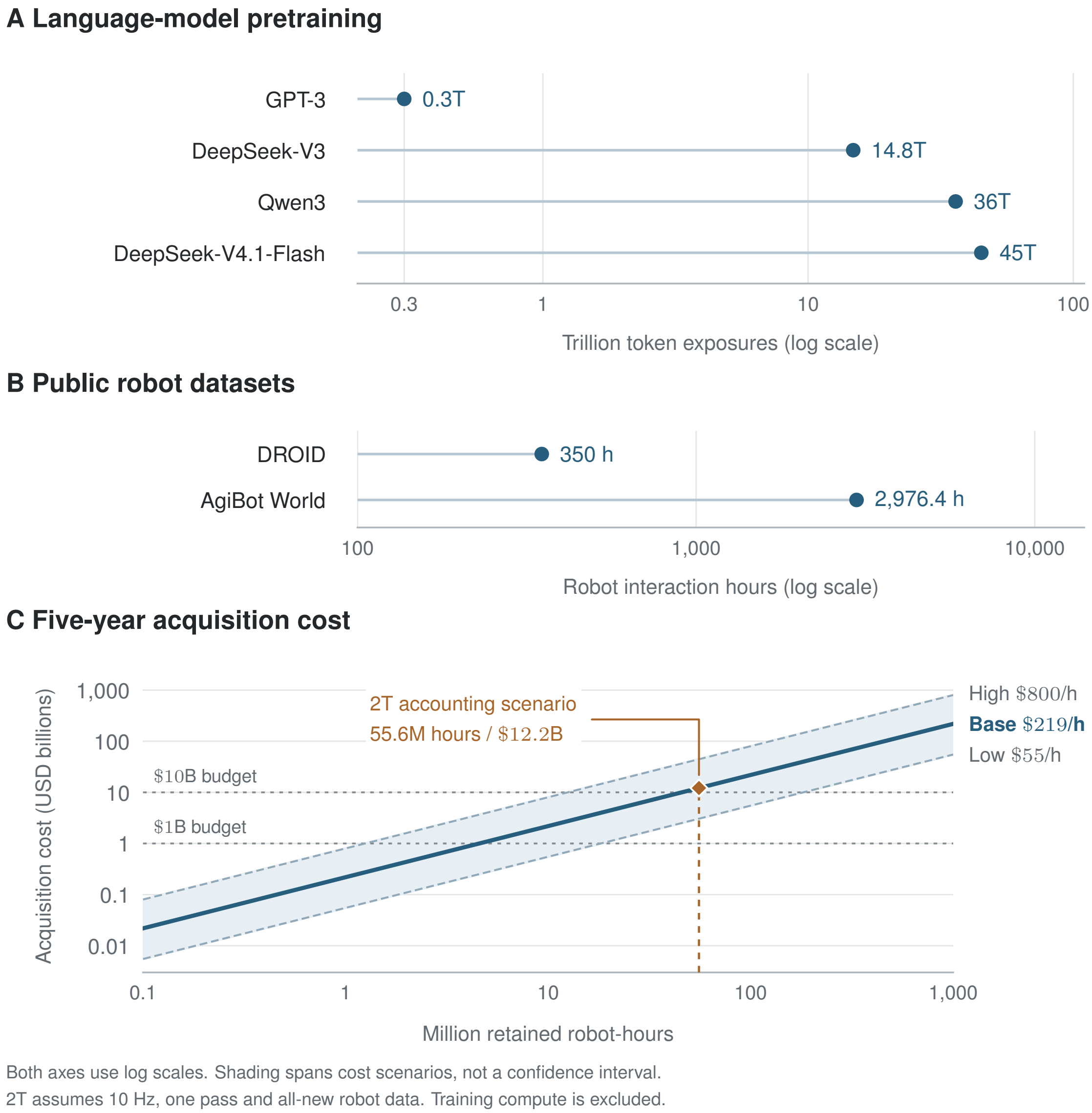

Figure 1: **Public data scales and illustrative acquisition cost. A:** Developer-reported model pretraining exposures, before unquantified post-training [1–4]. **B:** Reported robot interaction hours for selected public releases: DROID (350 hours) and AgiBot World (2,976.4 hours) [5, 6]. These totals are not unique examples or comparable measures of learning, and the releases are not an industry census. Dyna-2 separately reports over one million hours of human video; no robot-hour equivalence is assumed [7]. **C:** The solid blue line gives the base five-year estimate of $219 per retained hour. Pale shading spans the low and high cost scenarios of $55 and $800 per hour, with dashed boundaries; it is not a confidence interval. Dotted horizontal lines mark illustrative $1B and $10B budgets. The amber diamond marks the 2T accounting scenario: 55.6 million retained robot-hours and $12.2B at base cost, assuming 10 Hz, one pass and all-new robot data. This is not a capability threshold; model training compute is excluded.

change the inferred data requirement by $700/53 = 13.2$ times without adding physical experience. Accordingly, the report also prices hours directly.

Existing ecosystem. Reusable public datasets and internet-pretrained models carry zero incremental acquisition charge here. Their transfer benefit is allowed through α, but cannot be assigned an evidence-based universal discount. The main tables price an assumed *additional private corpus*. They are not a claim that every company must acquire that corpus.

2 Build the acquisition cost from paid time

All modeled costs are in constant 2026 US dollars. Assume a five-year program ($Y = 5$), one operator per active station and $A = 2{,}000$ staffed working hours per year (250 days × 8 hours), including setup and resets. Paid-hour labels below refer to this staffed time; paid-leave benefits are already in loaded compensation. Collection equipment is bought at years 0 and 3: two purchases per station ($n = 2$), with zero terminal salvage. Future wages, equipment prices and methods are held constant. There is no assumed future data-market expansion or algorithmic breakthrough.

Input and unit	**Low**	**Base**	**High**
Loaded operator compensation w ($/paid hour)	20	45	70
Other collection operations o ($/paid hour)	5	15	30
Installed collection station K ($)	32,000	100,000	200,000
Annual maintenance m (fraction of K)	10%	15%	20%
Recording share of paid time ρ	75%	50%	$33\frac{1}{3}$%
Accepted share of recorded time q	80%	80%	60%
Retained share of paid time $u = \rho q$	60%	40%	20%
Five-year retained hours per station YAu	6,000	4,000	2,000
Five-year cash cost per station ($)	330,000	875,000	1,600,000
Cost per retained hour c ($)	**55.00**	**218.75**	**800.00**

Evidence versus assumptions. Figure's current data-creator posting starts at $30/hour before additional compensation/benefits [10]. Using the BLS full-time private-sector wage share of 68.5% gives $30/0.685 = \$43.80$; the base rounds to $45 loaded compensation [11]. This is a proxy, not Figure's actual payroll cost. Low-case $20 assumes less expensive international recruitment; it is not a quoted US wage. Using $36 instead in the low case gives $81.67 per retained hour.

Mobile ALOHA reports $32,000 for its research system, including onboard power and compute [12]. The low case deliberately reuses this 2024 nominal figure as an optimistic 2026 assumption, without an inflation uplift; it is not a humanoid price quote. Base/high station costs, maintenance, three-year equipment life, overheads and retention rates are scenario assumptions, not observed Tesla/Figure accounts. They do not represent confidence intervals or equal-capability hardware.

Other operations o cover allocated supervision, setup logistics, facilities, quality assurance, labeling and data handling/storage during the program; operator benefits and equipment maintenance are not counted again. The $15 base allocation is not an independently measured unit cost. Failures, resets, repairs and setup reduce ρ; rejected or duplicate recordings reduce q.

For each station, retained output is YAu, so:

$$c = \frac{YA(w+o) + K(n+Ym)}{YAu}, \qquad S = \frac{H}{YAu}, \qquad C = ScYAu = Hc. \tag{2}$$

The base calculation is fully visible:

$$c = \frac{(5)(2{,}000)(45+15) + 100{,}000[2 + (5)(0.15)]}{(5)(2{,}000)(0.50)(0.80)} = \frac{875{,}000}{4{,}000} = \$218.75.$$

Scope. This is cash expenditure on acquisition, including replacement equipment. It excludes model development/training compute, general R&D, financing, tax, certification and deployment fleets. Instant fleet availability and linear overhead are simplifying assumptions; manufacturing and hiring ramps could extend the schedule. Zero salvage is conservative. These mixed effects mean the estimate is neither a universal floor nor a ceiling.

3 Program totals and sensitivity

	Five-year acquisition cost ($ billion)			Base stations
Retained experience	**Low**	**Base**	**High**	**and operator FTEs**
1 million hours	0.055	0.219	0.800	250
10 million hours	0.550	2.188	8.000	2,500
100 million hours	5.500	21.875	80.000	25,000
300B proxy: 8.333 million hours	0.458	1.823	6.667	2,084
2T proxy: 55.556 million hours	3.056	12.153	44.444	13,889

The first three rows are illustrative acquisition programs, not capability forecasts. Totals use unrounded hours and fractional fleet equivalents; physical station/headcount requirements are rounded up. At one shift per station, operator FTEs equal stations; support staff are budgeted in o.

For the two-trillion case, the working is:

$$H = \frac{2 \times 10^{12}}{3{,}600 \times 10} = 55{,}555{,}556 \text{ hours}, \quad S = \frac{55{,}555{,}556}{4{,}000} \simeq 13{,}889, \quad C = H \times 218.75 = \$12.153 \text{ billion}.$$

Its cost components are **$6.250B operator compensation + $2.083B other operations + $2.778B equipment purchases + $1.042B maintenance**. This makes the labor burden, rather than just the price of robots, explicit.

Change one assumption at a time. The following starts from the two-trillion base case; no savings are multiplied together.

Change from base	**$/retained hour**	**$ billion**
None	218.75	12.15
Three staffed shifts: $A = 6{,}000$	172.92	9.61
Retention $u = 60\%$ instead of 40%	145.83	8.10
Five training exposures per transition: $e = 5$	218.75	2.43
Only 10% residual real data: $\alpha = 0.1$	218.75	1.22
Accounting rate 2 Hz instead of 10 Hz	218.75	60.76

Three shifts amortize equipment more intensively; paid labor per retained hour is unchanged. This optimistic variant preserves equipment life and maintenance despite additional wear. Figure already advertises rotating 24/7 shifts [10], so one shift is not an industry maximum. The reuse and residual-data rows are arithmetic sensitivities: *neither demonstrates equivalent learning*. The α row prices only the remaining real-data component; newly acquired human/synthetic data would add cost. Changing accounting frequency cannot itself create capability.

An affordability condition, not an impossibility proof. At base costs, a $1B acquisition budget buys $10^9/218.75 = 4.57$ million retained hours. A $1B cap would therefore rule out the base 8.33M-hour program *under these assumptions*. But the companies' actual data budgets and minimum effective hours are not known. Tesla reported $43.524B in cash, equivalents and short-term investments at June 30, 2026; this is company liquidity, not a robotics allocation [13]. Figure announced more than $1B for data and compute combined over the year following August 2026 [14]. Neither figure proves a program is affordable or profitable. Economic feasibility additionally requires expected returns, time to revenue, competing capital needs and financing costs.

4 Earlier estimates and the strongest counterargument

Two closely matched blog estimates. Chris Paxton (June 2025) starts from two trillion tokens at 10 useful frames/second. This is $2 \times 10^{12}/(10 \times 3{,}600 \times 24 \times 365.25) = 6{,}338$ continuous robot-years; his utilization adjustment rounds the requirement to roughly 70,000 robot-years. He discusses a possible $1B 1,000-robot collection project and explicitly identifies Tesla and Figure as potential funders [8]. His token target and productivity penalty are thought-experiment inputs, not measured scaling requirements.

Rayan Malik (August 2025) assumes 1.1T tokens, 8 Hz, three tokens per transition and 600 retained hours per robot-year:

$$\frac{1.1 \times 10^{12}}{8 \times 3 \times 3{,}600 \times 600} = 21{,}219 \text{ robot-years}.$$

At 5,000 robots this takes 4.24 years. His \$550M initial equipment plus \$323M annual operations imply $0.550 + 0.323(4.244) = \$1.92\text{B}$ [15]. However, one supervisor per ten robots describes supervised autonomy, not one-to-one teleoperation. His parameter/token scaling assumption does not identify a robotics capability threshold. These estimates motivate billions in spending, but do not demonstrate single-company impossibility.

Current alternatives already weaken the all-teleoperation premise. These are present-day observations, not assumptions about future breakthroughs: Figure's Index reports 16M uploaded human videos and \$15M in creator payouts (August 2026), but does not disclose cumulative accepted hours, preventing a defensible dollar-per-retained-hour calculation [14]. Figure subsequently reports success improving from 9% to 56% on three behaviors across 30 unseen homes after Index pretraining, with robot adaptation data held fixed [16]. Those company-reported results concern environment generalization on selected behaviors, not arbitrary unseen tasks. Dyna-2 reports over one million human-video hours [7]. 1X reports 900 human-video hours and 70 hours of robot fine-tuning, alongside a separate inverse-dynamics model trained on 400 robot hours [17]. Do not omit that additional robot-data dependency or assume those datasets are disjoint. None establishes a universal human-to-robot exchange rate.

Where the human harness argument becomes testable. Let η be target-specific robot-equivalent learning value per retained human hour, c_H the fully loaded human-hour cost, and s the fraction of required effective experience replaced. A simple mixture model is:

$$C_{\text{mix}} = H\left[(1-s)c + s\frac{c_H}{\eta}\right] + C_{\text{transfer}}.$$

Ignoring transfer overhead only for the marginal comparison, human collection is cheaper if $\eta > c_H/c$. For an *illustrative, unmeasured* $c_H = \$20$, the base threshold is $\eta > 20/218.75 = 0.0914$. Thus even one-tenth the learning value could be competitive—if experimentally demonstrated and if transfer costs are small. Wearables record human motion and observations; they do not directly record the target robot's joint state, actuator commands or contact dynamics. Calling this robot proprioception would overstate what is acquired.

Defensible conclusion. Paying for tens of millions of robot-hours through dedicated operators can demand a multibillion-dollar program. That supports investigating human collection, autonomous practice, simulation and existing pretrained models. It does not establish that one company cannot reach general robotics. The pivotal missing evidence is the relationship between diverse retained experience and a predefined held-out capability benchmark, followed by measured retention and transfer costs. Counting tokens alone cannot supply that evidence.

References

[1] T. B. Brown, B. Mann, N. Ryder, et al. *Language Models are Few-Shot Learners*. Tables 2.1–2.2: training exposures and component corpus sizes. 2020. arXiv: 2005.14165.

[2] DeepSeek-AI. *DeepSeek-V3 Technical Report*. Abstract: pretraining token count, before supervised fine-tuning and reinforcement learning. Dec. 27, 2024. arXiv: 2412.19437.

[3] Qwen Team. *Qwen3 Technical Report*. Section 3: pretraining tokens; a historical comparator, not an estimate for later releases. May 2025. arXiv: 2505.09388.

[4] DeepSeek-AI. *DeepSeek-V4.1-Flash: Pushing the Limits of KV Cache Compression*. Section 4.2.2: multimodal pretraining tokens. Sept. 17, 2026. arXiv: 2609.19969.

[5] A. Khazatsky, K. Pertsch, S. Nair, et al. *DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset*. Abstract: 350 hours of robot interaction data. Mar. 19, 2024. arXiv: 2403.12945.

[6] Team AgiBot-World. *AgiBot World Colosseo: A Large-scale Manipulation Platform for Scalable and Intelligent Embodied Systems*. Version 4, Section III: 2,976.4 hours; historical release. Aug. 4, 2025. arXiv: 2503.06669.

[7] Dyna Robotics. *Dyna-2*. Company research report on million-hour human-video pretraining. Aug. 2026. (Visited on 09/28/2026).

[8] C. Paxton. *How Can We Get Enough Data to Train a Robot GPT?* Original blog thought experiment, rather than empirical cost accounts. Substack. June 10, 2025. (Visited on 09/28/2026).

[9] K. Pertsch et al. *FAST: Efficient Action Tokenization for Vision-Language-Action Models*. Table I: token counts for one-second action chunks. Jan. 16, 2025. arXiv: 2501.09747.

[10] Figure. *Helix Data Creator. San Jose, CA*. Official job posting; base wage and rotating-shift schedule. (Visited on 09/28/2026).

[11] US Bureau of Labor Statistics. *Compensation Costs for Private Industry Workers Averaged $46.89 per Hour Worked in June 2026*. Full-time wages as a share of total employer compensation. The Economics Daily. Sept. 17, 2026. (Visited on 09/28/2026).

[12] Z. Fu, T. Z. Zhao, and C. Finn. *Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation*. Page 1: research hardware cost. Jan. 4, 2024. arXiv: `2401.02117`.

[13] Tesla. *Form 10-Q for the Quarter Ended 30 June 2026*. Note 2: cash and short-term investments. US Securities and Exchange Commission. 2026. (Visited on 09/28/2026).

[14] Figure. *Introducing Index*. Company-reported videos, creator payouts and planned data/compute spending. Aug. 25, 2026. (Visited on 09/28/2026).

[15] R. Malik. *Scaling to a Trillion Tokens in Robotics*. Original blog estimate of a token target and fleet/operating costs. Substack. Aug. 20, 2025. (Visited on 09/28/2026).

[16] Figure. *Helix 2.5: Zero-Shot 30-Home Generalization*. Company research report; controlled comparison on three behaviors. Sept. 17, 2026. (Visited on 09/28/2026).

[17] 1X. *World Model: Self-Learning*. Company research report; human-video, robot-adaptation and inverse-dynamics inputs. Jan. 12, 2026. (Visited on 09/28/2026).

Method and verification. Sources were reviewed through 28 September 2026. Research and company reports establish only their stated experiments; blogs supply useful estimation methods, not validated thresholds. Every uncited cost input is an explicit scenario assumption. All totals are recomputed from equations (1) and (2); displayed rounding is not an uncertainty estimate. No future cost decline, free robot-equivalent data or compounded transfer multiplier is assumed in the main totals.