

How Many Trials Back a Force-Aware Manipulation Claim?

A Statistical Audit of 2026 Real-Robot Evaluations

Clement Aidoo
CANLABS, Ghana
`clement.aidoo@canlabs.org`

October 2026 (version 2)

Abstract

Papers on manipulation with force, torque, tactile or motor-current feedback usually support their claims with real-robot success rates measured over a handful of trials. We audited how much statistical evidence those numbers carry. Following a protocol frozen before any data were extracted, we screened 89 arXiv papers retrieved by six fixed queries (first posted January–September 2026), included 58, and extracted 377 binary comparisons of the proposed method against its strongest baseline or against its own ablation without the force/touch signal (337 in the primary set). A second, blinded extraction of every paper, by separate instances of the AI assistant used for the first, checked the data before analysis, and the author re-coded a seeded random 20% of papers by hand (19 of 19 inclusion decisions and 100% of 88 binary rows matched). The median condition was evaluated on 20 trials (interquartile range 15–30); 17% of papers reported any interval or error bar on success, 5% any significance test and 5% the number of training seeds. The proposed method scored higher in 91% of comparisons, but the 95% interval for the difference excluded zero in only 53% (49% against the strongest baseline). After a Holm correction within each paper, 23% of papers had every headline comparison resolved. At the reported trial counts the median comparison could detect, with 80% power, only a gap of about 34 percentage points; only 6% of comparisons had 80% power to detect a true 20-point improvement (median power 0.19). In simulation, two equally good methods compared on 10 trials each show an apparent gain of 10 points or more 41% of the time. None of this says the methods do not work: most reported gaps are large and in the expected direction. It says that the field’s typical evaluation cannot distinguish a real 20-point improvement from noise, and we give a short checklist and a sequential-testing illustration that would cost few additional trials. All data, code and the frozen protocol are released.

1 Introduction

Learning-based manipulation increasingly uses contact signals: wrist force/torque sensors, fingertip tactile arrays, vision-based tactile sensors, and actuator effort read from motor current or joint torque. The main evidence that such a signal helps is almost always a table of real-robot success rates, comparing the proposed method with baselines and with the same method without the signal. Real-robot trials are slow and require manual resets, so these tables are built from small numbers of trials, typically 10 or 20 per cell.

Small samples are not a flaw by themselves. The issue is what a reader can conclude from them. Two methods with the same true success rate of 50% will, over 10 trials each, differ by at least 10 points about 41% of the time. Kress-Gazit et al. argued that robot learning should report trial counts, success criteria and statistical analysis rather than bare success rates [13], and

Snyder et al. showed that rigorous policy comparison often needs far more trials than the field uses, proposing sequential tests that stop early when the evidence is clear [16, 17]. Similar concerns led to recommended practice in deep reinforcement learning [9, 1, 6]. What has not been measured is how the current literature in one active area actually stands.

This paper does that for force-, tactile- and current-aware manipulation, an area that grew quickly in 2026 and in which our own work on motor-current sensing sits [2]. We ask four questions, fixed in advance: how many trials per condition are reported; how often uncertainty and tests are reported; how many headline comparisons remain distinguishable from zero under standard binomial statistics; and what size of improvement the reported trial counts could reliably detect. We add two illustrations: a simulation of how often equal methods look different by chance under common reporting practices, and the trial cost of a simple sequential test.

The audit is about evaluation practice, not about any paper’s merit. A comparison that is not statistically resolved is not evidence that a method does not work; in most such cases the reported gap is large and the honest conclusion is “promising, not yet established”. We therefore report results in aggregate, do not rank or name papers in the text, and release every extracted number with its source so that any reader, including the authors, can check and correct it.

Our contributions are:

1. A pre-specified, timestamped audit protocol and a dataset of 377 real-robot comparisons from 58 2026 papers, double-extracted, adjudicated and partly re-coded by hand (Sections 3, 4).
2. Measurements of trial counts, reporting practice, statistical resolution, power and required sample sizes for this literature (Section 5).
3. A practical note for planning: at the small n used in robotics, the standard arcsine sample-size formula undershoots the exact Fisher requirement (exact power 0.71–0.78 rather than 0.80 at its recommended n on a standard design grid). This is a known property of the normal approximation, but one that matters at these sample sizes (Section 5.4).
4. A short evaluation checklist and a sequential-testing illustration (Section 6).

2 Related Work

Evaluation practice in robot learning. Kress-Gazit et al. recommend explicit success criteria, reporting of trial counts and initial conditions, statistical analysis, and failure descriptions [13]. Snyder et al. formalize two-policy comparison as a composite hypothesis test and propose STEP, a near-optimal sequential test for binary success [16], later extended to partial-credit and continuous metrics with anytime-valid guarantees [17]. They note that a gap of 8–9 points between two real policies would need around 500 trials per policy to establish [16]. Our work is complementary: rather than proposing a test, we measure what the current literature in one area reports and how it fares under standard tests.

Statistical practice in machine learning. Henderson et al. showed that deep RL results vary strongly with seeds and implementation details [9]; Colas et al. discussed how many seeds a comparison needs [6]; Agarwal et al. recommended interval estimates and aggregate metrics that are robust with few runs [1]. Our setting differs in that the unit of evidence is a physical trial with a binary outcome, and the dominant source of uncertainty is binomial sampling of a handful of trials rather than seed variance.

Power and selective reporting. Low statistical power not only makes true effects hard to detect; it also inflates the size of effects that do pass a significance filter and lowers the probability

that a significant result is true [4, 12]. Gelman and Carlin call these Type M (magnitude) and Type S (sign) errors [8]. Selecting the best of several variants or tasks before reporting adds a further upward bias, which our simulation (A6) quantifies for the settings common in this literature.

3 Methods

The protocol was written and committed to a git repository before any success rate was extracted. Every later change is recorded as a dated deviation with a reason, and every deviation was committed before it was applied (Appendix A).

3.1 Search and screening

We ran six fixed queries on arXiv through the alphaXiv search interface (e.g. “force-aware manipulation policy real robot”, “motor current proprioceptive contact manipulation low-cost robot”; full list in the protocol) and screened every hit, plus twelve seed papers named in advance. Because the search tool ranks recent work first, nearly all hits were from 2026, so we narrowed the window to papers first posted between 2026-01-01 and 2026-09-30 and describe the result as a snapshot of 2026 work retrieved by fixed queries, not a census (deviation D1).

A paper was included if (I1) it evaluates a manipulation policy, controller or learning pipeline whose input, as deployed on the real robot, includes force, torque, tactile or actuator-effort signals; (I2) it reports a real-robot evaluation with a binary or binarisable success outcome; and (I3) trial counts are reported or recoverable for at least one comparison with a baseline or ablation. Camera-measured contact deformation counts as tactile (D4); signals used only as a training reward do not satisfy I1 (D6). Surveys, sensor-hardware papers and simulation-only studies were excluded.

3.2 Extraction

For every real-robot task and condition we extracted the number of trials and successes of the proposed method (P) and of two kinds of comparator (X): *kind B*, the strongest baseline by reported rate, restricted to external or author-built baselines (not ablations that still use the signal, not oracle variants, not baselines containing the proposed method’s own modules; D5); and *kind A*, the proposed method with the force/touch input removed, or a baseline the paper itself identifies as such (D2b). When only a percentage was given, successes were $\text{round}(\text{rate} \times n)$ and the row was flagged if the percentage was not k/n for an integer k . Tasks scored with partial credit were recorded and counted in the trial-count and reporting statistics but excluded from binomial tests (D2a). We also recorded whether each paper reported intervals or error bars on success, any significance test, the number of training seeds, an explicit success criterion, and randomised or matched initial conditions. Text–table inconsistencies were recorded neutrally.

Where a paper reports several variants of its method, or several rows for the same task and comparator kind, the primary analysis keeps the first-listed row, which follows the paper’s own table order; all rows are analysed as a sensitivity check (D7a).

3.3 Independent verification

After the first extraction was complete and before results were finalised, every paper was re-extracted by independent extractors (separate instances of the same language-model assistant used for the first extraction, given only the protocol rules and the papers). Verifiers first screened all 89 papers against I1–I3 from the papers alone and wrote their decisions before seeing anything

else. They then received, for each included paper, only the task names, comparator names and comparison kinds of the first extraction, with all trial counts, successes and decisions removed, filled in the numbers from the source, and judged whether each comparator met the kind-A and kind-B rules. Every disagreement was adjudicated against the source paper, and corrections were applied by script, each with a logged reason.

After the first version of this paper was posted, the author re-coded a random sample by hand (seed 20261016; 20% of included and 20% of excluded papers, 19 papers in all) without consulting the extraction records. Coding sheets listed only paper links, and, for included papers, task, comparator and comparison-kind names; screening was completed before the row sheet was opened. Filled sheets were locked by hash before scoring, so the agreement reported is the pre-resolution one.

3.4 Analyses

All analyses were fixed in the protocol (A1–A7) with implementation details fixed before the first run (D7).

A1–A2 Distribution of trials per condition, and reporting rates.

A3 For each binary comparison: Wilson 95% intervals per arm [19]; the Newcombe hybrid-score 95% interval for the difference (method 10) [14]; the two-sided Fisher exact test [7], with Boschloo’s test as a sensitivity analysis [3]. A comparison is *resolved* if the difference interval excludes zero.

A4 Per paper: whether at least one, a majority, or all of its comparisons are significant at $\alpha = 0.05$, unadjusted and after Holm correction within the paper [11].

A5 For each comparison, the minimum detectable difference at the reported n given the observed comparator rate, and the exact power of the two-sided Fisher test ($\alpha = 0.05$, equal arms of the smaller reported n , computed by full enumeration of the 2×2 tables) to detect pre-set gains of 15, 20 and 30 points over that rate. Power is deliberately *not* evaluated at each comparison’s own observed gap, since such post hoc power is a restatement of the comparison’s p -value [10] (deviation D9). Separately, on a fixed design grid we compare the sample size recommended by the arcsine approximation [5] with the exact Fisher requirement.

A6 Null simulation: both methods share a true rate $p \in \{0.3, 0.5, 0.7\}$; a study runs $n \in \{5, 10, 20, 50\}$ trials on each of $k \in \{1, 3, 5\}$ tasks and reports the best of $m \in \{1, 3, 5\}$ variants of its method against one baseline. We record how often the mean apparent gain across tasks is at least 10 or 20 points (200,000 replicates).

A7 A Wald sequential probability ratio test [18] on discordant pairs of alternating trials, designed for 50% vs 70% (one-sided $\alpha = 0.025$, power 0.8), compared with the fixed-sample exact test.

Code uses SciPy [15] and is unit-tested against published values (e.g. Newcombe’s worked examples) and against Monte Carlo estimates.

4 Verification Results

The two screenings agreed on 81 of 89 inclusion decisions (Cohen’s $\kappa = 0.80$). For the 437 comparison rows checked, the blinded re-extraction reproduced all four numbers exactly in 99.5% of binary rows and agreed on the outcome type (binary vs partial credit) in 98%. Verifiers questioned the

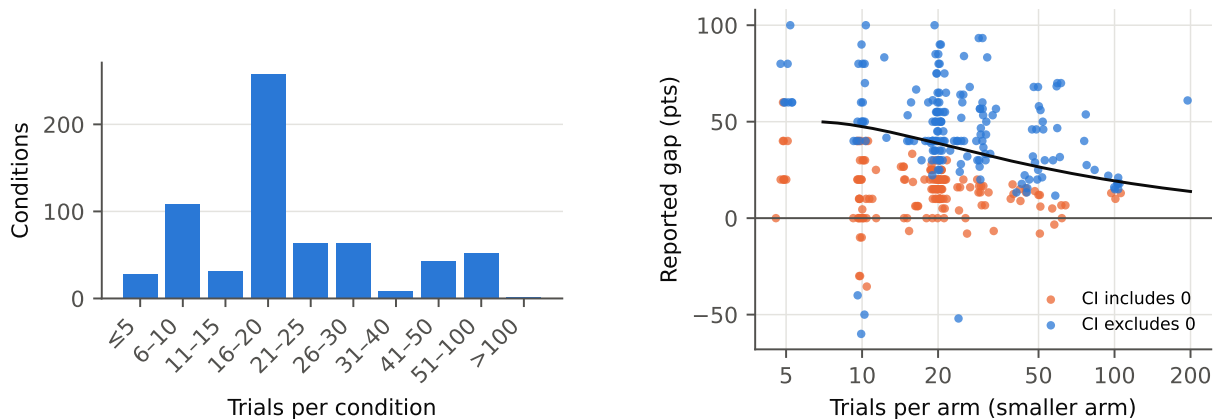


Figure 1: Left: trials per evaluated condition (659 conditions, 58 papers). **Right:** reported gap between the proposed method and its comparator against the number of trials in the smaller arm, for the 337 primary binary comparisons (log scale, small horizontal jitter). Blue: Newcombe 95% interval excludes zero; orange: it does not. The black line is the gap an equal-arm study can detect with 80% power when the comparator succeeds 50% of the time.

comparator choice in 30 kind-B rows and 9 kind-A rows. After adjudication, 60 rows were corrected, added or removed across 16 papers, and 6 inclusion decisions changed. The human re-check agreed on 19 of 19 inclusion decisions ($\kappa = 1.00$) and on 56 of 60 reporting-field codes for the 12 papers both coded as included; all four reporting disagreements, three of them on whether success was explicitly defined, were resolved in favour of the extraction on re-reading. Of 96 comparison rows, the hand coding reproduced all four numbers exactly in 100% of the 88 binary rows and agreed on every outcome type. It questioned 3 comparator choices (all on a partial-credit task, so no test was affected) and found 4 comparisons missing from the extraction, two of which had extractable counts and were added. All results below use the adjudicated data with these corrections; both logs are released.

5 Results

5.1 Trial counts and reporting (A1, A2)

Across 659 evaluated conditions in the 58 included papers, the median number of trials was 20 (interquartile range 15–30; Figure 1). 21% of conditions used 10 or fewer trials, 65% used 20 or fewer, and 12% used 50 or more.

Table 1 summarises reporting. Most papers defined success explicitly (76%); 36% clearly described randomised or matched initial conditions. Few reported any uncertainty: 17% gave an interval or error bar on success, 5% a significance test, and 5% the number of training seeds. Several papers that did report intervals or tests used them well, for example by stating explicitly that a gain was within seed variation or not significant; these practices are noted in the released data.

5.2 How many comparisons are resolved? (A3)

Table 2 gives the main result. The proposed method scored higher than its comparator in 91% of the 337 primary comparisons, with a median gap of 28 points. The Newcombe interval for the difference excluded zero in 53% of comparisons: 49% against the strongest baseline and 57%

Table 1: Reporting practice in the 58 included papers (real-robot results).

Item reported	Share of papers
Explicit success criterion	76%
Randomised or matched initial conditions	36%
Interval or error bars on success rate	17%
Number of training seeds	5%
Any significance test	5%

Table 2: Statistical resolution of the primary binary comparisons. “Resolved” = Newcombe 95% interval for the difference excludes zero. MDD: minimum detectable improvement over the observed comparator rate at the reported n , 80% power (arcsine approximation). Power rows: exact two-sided Fisher power, equal arms of the smaller reported n , to detect the stated gain over the observed comparator rate.

	All	vs strongest baseline (B)	vs no-signal ablation (A)
Comparisons	337	176	161
Proposed method higher	91%	93%	88%
Median reported gap (pts)	28	22	32
Resolved (CI excludes 0)	53%	49%	57%
Fisher $p < 0.05$	44%	42%	47%
Boschloo $p < 0.05$	50%	47%	55%
Median MDD (pts)	34	33	37
Power ≥ 0.8 for +15 pts	1%	2%	0%
Power ≥ 0.8 for +20 pts	6%	8%	4%
Power ≥ 0.8 for +30 pts	25%	27%	23%
Median power, +20 pts	0.19	0.22	0.18

against the no-signal ablation. The Fisher test was significant in 44%; Boschloo’s more powerful unconditional test in 50%, with the two disagreeing on 21 comparisons. Using all 377 rows instead of the primary set gives 51% resolved, and excluding rows whose counts were reconstructed from ambiguous percentages gives 51%.

Figure 1 (right) shows why. Most comparisons sit at 10–30 trials per arm, where a gap must be roughly 30–50 points to be detected reliably. Many reported gaps of 20–40 points fall below that line: large enough to be interesting, too small to be distinguished from noise at that sample size.

5.3 Paper-level resolution (A4)

Papers usually make several comparisons, and their claims rest on the pattern. Without correction, 81% of the 52 papers with binary comparisons had at least one significant comparison, 37% had a majority, and 23% had all. After Holm correction within each paper these fell to 73%, 29% and 23%. Against the strongest baseline alone, 22% of papers had every comparison resolved after correction; against the no-signal ablation, 22%. The claim that the force or touch signal itself is responsible for the gain therefore rests, in most papers, on at least one comparison that the data cannot yet distinguish from chance.

Table 3: Planning a two-arm comparison for 80% power (two-sided $\alpha = 0.05$): trials per arm recommended by the arcsine approximation, the exact power of the Fisher test at that n , and the exact required n .

Comparator rate	Target gain	Arcsine n	Exact power at arcsine n	Exact n
0.1	+15 pts	97	0.75	110
0.1	+20 pts	60	0.73	69
0.1	+30 pts	30	0.71	36
0.3	+15 pts	162	0.76	174
0.3	+20 pts	93	0.75	102
0.3	+30 pts	42	0.76	48
0.5	+15 pts	170	0.78	183
0.5	+20 pts	93	0.75	102
0.5	+30 pts	38	0.73	44
0.7	+15 pts	119	0.76	131
0.7	+20 pts	60	0.73	69

5.4 What could these evaluations detect? (A5)

At the reported trial counts, the median comparison could detect with 80% power only an improvement of about 34 points over its comparator (Table 2). Put the other way, if a method truly improved on its comparator by 20 points, the evaluations in this literature would have had a median power of 0.19 to show it under an exact test, and only 6% of comparisons had the conventional 80%. For a true 30-point improvement the median power was 0.49, and 25% of comparisons reached 80%. Many of the gaps that *were* resolved are therefore likely to be larger than the true effects [4, 8].

Planning more trials requires a sample-size formula, and the usual one is the arcsine (normal) approximation. Table 3 compares it with the exact Fisher requirement on a fixed grid of comparator rates and target gains. At the n the formula recommends, the exact test has power 0.71–0.78 (median 0.75) rather than 0.80, and the exact requirement is a median 1.13 and at most 1.20 times the formula’s value. The shortfall is a known property of the approximation with a conservative test, and it grows when one arm is expected near 0% or 100%; we note it because a robotics group that plans with the formula will be underpowered by design. Exact power for two binomial arms takes a few lines of code (released).

5.5 How often do equal methods look different? (A6)

Table 4 and Figure 2 show the null simulation at $p = 0.5$ (results at 0.3 and 0.7 are similar and released). With one task and 10 trials per arm, two equally good methods differ by at least 10 points 41% of the time; at 20 trials, 32%. Reporting the best of several variants raises these numbers: at 20 trials, the best of five variants shows a gain of at least 20 points 36% of the time. Averaging over more tasks reduces the noise of the mean gain, but selection across variants largely undoes this: at 10 trials per arm, reporting the best of three variants over three tasks still shows a mean gain of 10 points or more 47% of the time. Non-monotonicity at very small n (5 vs 10 trials) is a discreteness effect: with 5 trials the smallest possible nonzero gap is 20 points.

5.6 What would a sequential test cost? (A7)

Table 5 shows a simple sequential test that alternates trials of the two methods and stops when the evidence crosses a boundary. Designed to detect 70% vs 50%, it needed on average 67 trials

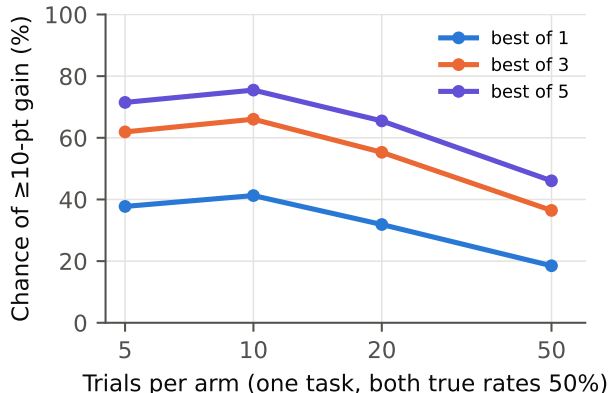


Figure 2: Chance that a method with no true advantage shows an apparent gain of at least 10 points over a baseline, both with true success rate 50%, one task, when the best of m variants of the method is reported (200,000 simulated studies per point).

per method when that was the truth, against 102 for a fixed-sample exact test with the same error rates, and stopped after 39 trials on average when the methods were equal. Much larger true gaps, which several papers report, are confirmed in fewer trials still. Near-optimal sequential tests for robot policy comparison, with proper handling of composite hypotheses and a maximum budget, are available [16, 17]; the SPRT here is only an illustration of the order of magnitude.

6 Discussion

What the numbers mean. The pattern is consistent. Proposed methods almost always score higher, and the gaps are often large, but the typical evaluation of 10–30 trials per arm can only reliably distinguish gaps of roughly 30–50 points from noise. About half of individual comparisons, and a minority of papers taken as a whole, are statistically resolved, and a true 20-point improvement would usually go undetected. Low power also means that the gaps which do pass a significance filter are likely to overstate the true effect [4, 8], and our null simulation shows that common reporting choices, such as reporting the best variant or configuration, make 10–20 point gains easy to obtain by chance. None of this implies that the force or touch signals do not help; the direction of the evidence is clearly positive. It implies that the size of the benefit, and which design choices produce it, are much less certain than the tables suggest.

Why it happens. Real-robot trials are expensive, and the incentive is to spend them across many tasks and baselines rather than on many trials per comparison. A paper with six tasks, four baselines and an ablation at 20 trials each runs 720 trials, a substantial effort, yet each cell is underpowered. Statistical reporting is also not expected by reviewers in this area. The fix is therefore partly cultural and partly a matter of allocating the same effort differently.

A short checklist. Based on the audit, we suggest the following for real-robot comparisons, most of which cost no extra trials:

1. Report successes *and* trials (k/n) for every cell, not only percentages.
2. Give a 95% interval for each rate (Wilson) and for each headline difference (Newcombe).
3. Pre-specify one or two headline comparisons per paper, test them (Fisher or Boschloo), and correct for multiplicity when several comparisons support one claim.

Table 4: Null simulation (A6), true rate 0.5 for both methods: probability (%) that the reported mean gain over k tasks, taking the best of m variants, is at least 10 or 20 points.

n	k	Pr(gain ≥ 10)			Pr(gain ≥ 20)		
		$m=1$	$m=3$	$m=5$	$m=1$	$m=3$	$m=5$
5	1	38	62	71	38	62	71
5	3	29	52	62	18	36	45
5	5	24	45	55	10	22	30
10	1	41	66	75	25	46	56
10	3	26	47	58	8	18	24
10	5	18	36	46	3	7	10
20	1	32	55	65	13	28	36
20	3	16	32	41	2	5	7
20	5	9	20	27	0	1	1
50	1	18	36	46	3	7	10
50	3	5	11	16	0	0	0
50	5	1	4	5	0	0	0

Table 5: Wald SPRT on discordant pairs, designed for $p_X = 0.5$ vs $p_P = 0.7$ (one-sided $\alpha = 0.025$, power 0.8, at most 500 pairs; 20,000 replicates). A fixed-sample two-sided Fisher test with 80% power at 0.5 vs 0.7 needs 102 trials per arm.

True rates (p_X, p_P)	Mean trials per arm	Pr(decide for P)	Pr(no decision)
(0.5, 0.5)	39	0.024	0.000
(0.5, 0.6)	76	0.321	0.001
(0.5, 0.7)	67	0.844	0.000
(0.5, 0.8)	42	0.983	0.000
(0.5, 0.9)	29	1.000	0.000

4. Plan n with exact binomial power for the gap you expect, not the normal approximation; when the budget is fixed, report the minimum detectable gap.
5. Prefer fewer, better-powered comparisons for the central claim; pool clearly related tasks only when the claim is about the pool.
6. Use a sequential test when trials are expensive and gaps may be large [16].
7. State how the reported variant or configuration was chosen, and report the others.
8. Say “not resolved at this sample size” rather than “outperforms” when the interval includes zero.

7 Limitations

Sample. The audit covers papers retrieved by six fixed queries from one search interface that ranks recent work first; it is a snapshot of 2026 work, not a census, and may over-represent prominent or recent papers. Most are preprints, and published versions may differ. Whether the same pattern

holds for papers accepted at refereed robotics venues is an open question that an audit of, for example, conference proceedings could answer.

Extraction. Both extractions were done by the same kind of language-model assistant under the author’s supervision; although the second was blinded to values and decisions, the two may share systematic errors, and verifiers received the first extraction’s task and comparator list, so omissions in that list were only partly checked (verifiers were asked to report missing comparisons). Agreement between two instances of the same model is not equivalent to agreement between independent human coders; the hand re-check of a random 20% sample (Section 4) addresses this, but it was done by the author, who designed the protocol and had seen the aggregate results, so it tests extraction accuracy rather than offering a fully independent second opinion, and the remaining 80% of papers were checked only by the model. We will invite each paper’s authors to confirm their extracted rows and report the outcome in a revision. Coding rules required judgment (which baseline is eligible as “strongest”, what counts as a no-signal ablation), and we made several of these rules explicit only during extraction; all are logged with dates. One set of rules (D8) was written during verification, after a first run of the analysis code on the unverified data whose results were not used; it was prompted by the verifiers’ screening and is applied uniformly. Reconstructing counts from percentages introduces small errors, which the sensitivity analysis addresses.

Statistics. We treat trials as independent Bernoulli outcomes and the two arms as independent. Many papers use matched initial conditions, for which paired analyses would be more powerful, and some pool trials across training seeds, which ignores clustering. We test each comparison on its own, whereas papers often argue from consistency across tasks; a hierarchical model would use that information, at the cost of further assumptions. Partial-credit outcomes were excluded from tests. These assumptions cut both ways: ignoring pairing makes our tests conservative, while ignoring clustering, operator effects and drift across trials makes them anti-conservative, so the resolution rates above should be read as approximate rather than as biased in a known direction.

Scope of conclusions. Unresolved comparisons are not evidence of no effect. We did not attempt to judge any paper’s contribution, and the per-paper data are released for checking, not ranking.

8 Conclusion

In 2026 work on force-, tactile- and current-aware manipulation, real-robot claims rest on a median of 20 trials per condition, rarely with any statement of uncertainty. Under standard binomial statistics about half of individual comparisons, and a small minority of papers as a whole, are resolved, and the median comparison could only reliably detect a gap of about 34 points; a true 20-point improvement would have been detected with 80% power in only 6% of comparisons. The methods may well be as good as reported; the evaluations mostly cannot tell. Reporting k/n with intervals, testing a few pre-specified comparisons with exact methods, and allocating trials to the central claim, or using sequential tests, would make the same effort far more informative.

Data and code availability

The frozen protocol with all dated deviations, per-paper extraction records with sources and notes, the blinded verification packets and outputs, the adjudication log, the analysis code with unit tests, and all generated tables are archived on Zenodo at doi:10.5281/zenodo.23452145 (version 2.0; version 1.0 is doi:10.5281/zenodo.23442248), including the full git history, so the protocol

timestamps can be checked. Every number in this paper is read from files generated by the released scripts.

Acknowledgements and disclosure

Extraction, verification, adjudication and analysis code were carried out with the assistance of an AI model (Claude, Anthropic) under the author’s direction; every decision is logged in the released repository, and the author takes responsibility for the content.

References

- [1] R. Agarwal, M. Schwarzer, P. S. Castro, A. C. Courville, and M. G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [2] C. Aidoo. Gravity pivoting from motor current: calibration-free angle estimation and the sensing budget it requires. *engrXiv preprint*, doi:10.31224/8492, 2026.
- [3] R. D. Boschloo. Raised conditional level of significance for the 2×2 -table when testing the equality of two probabilities. *Statistica Neerlandica*, 24(1):1–9, 1970.
- [4] K. S. Button, J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, 2013.
- [5] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum, 1988.
- [6] C. Colas, O. Sigaud, and P.-Y. Oudeyer. How many random seeds? Statistical power analysis in deep reinforcement learning experiments. *arXiv:1806.08295*, 2018.
- [7] R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.
- [8] A. Gelman and J. Carlin. Beyond power calculations: assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651, 2014.
- [9] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.
- [10] J. M. Hoenig and D. M. Heisey. The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1):19–24, 2001.
- [11] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.
- [12] J. P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005.
- [13] H. Kress-Gazit, K. Hashimoto, N. Kuppuswamy, P. Shah, P. Horgan, G. Richardson, S. Feng, and B. Burchfiel. Robot learning as an empirical science: best practices for policy evaluation. *arXiv:2409.09491*, 2024.
- [14] R. G. Newcombe. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, 17(8):873–890, 1998.
- [15] P. Virtanen et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020.

- [16] D. Snyder, A. J. Hancock, A. Badithela, E. Dixon, P. Miller, R. A. Ambrus, A. Majumdar, M. Itkina, and H. Nishimura. Is your imitation learning policy better than mine? Policy comparison with near-optimal stopping. arXiv:2503.10966, 2025.
- [17] D. Snyder, A. Badithela, N. Matni, G. Pappas, A. Majumdar, M. Itkina, and H. Nishimura. Beyond binary success: sample-efficient and statistically rigorous robot policy comparison. arXiv:2603.13616, 2026.
- [18] A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.
- [19] E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

A Protocol deviations

All deviations were committed to the repository, with timestamps, before they were applied. D1–D7 were made before any analysis was run; D8 was made during verification, after a first run on unverified data; D9 was made after the draft results were known; the hand re-check (D10) followed the first posted version. In brief: **D1** narrowed the window to 2026-01-01 – 2026-09-30 because the search tool returns mostly recent work. **D2** (a) excluded partial-credit outcomes from binomial tests; (b) defined kind A as the proposed method with the signal removed. **D3** kept eligible papers with no stated trial counts in the reporting statistics. **D4** counted camera-measured contact deformation, and RGB-generated tactile signals consumed through a tactile channel, as tactile. **D5** restricted kind B to external or author-built baselines. **D6** required the signal to be an input at deployment. **D7** fixed analysis details before the first run: primary-row selection, Holm families, the reading of “best of m ”, exact power by enumeration, and the sequential-test design. **D8** (during verification) treated papers with only partial-credit outcomes as failing I2, admitted per-trial binary outcomes not labelled “success”, and coded unresolved disagreements on two judgement fields conservatively. **D9** (after results, following internal review) replaced the planned A5 summary, power at each comparison’s observed rates, which restates its p -value, with exact power for pre-set gains and a fixed design grid; the observed-rate quantities remain in the released data and are not interpreted. **D10** (after version 1) added the author’s blinded hand re-check of a seeded 20% sample; its corrections (three comparator names on a partial-credit task, two added comparisons) are logged. The full text of all deviations is in the released protocol.

B Screened papers

Table 6: All screened papers. Binary = primary binary comparisons; resolved = Fisher $p < 0.05$ after Holm correction within the paper. Per-row data, sources and coder notes are in the released data.

arXiv ID	Signal	Binary	Resolved	Median n	Status
2502.17432	force	–	–	–	excluded: date (D1)
2508.07118	tactile	–	–	–	excluded: date (D1)
2602.22088	force	16	1	10	included
2603.04531	tactile	–	–	–	excluded: I2
2603.05687	tactile	2	0	15	included
2603.08342	force	5	0	20	included

arXiv ID	Signal	Binary	Resolved	Median n	Status
2603.13869	tactile	0	0	–	included
2605.28412	mixed	–	–	–	excluded: I1
2606.04825	tactile	16	1	10	included
2606.06281	tactile	10	5	20	included
2606.08555	force	8	1	20	included
2606.09337	tactile	4	4	30	included
2606.10818	torque	4	2	25	included
2606.11184	mixed	4	2	20	included
2606.11767	tactile	0	0	100	included
2606.12406	current	–	–	–	excluded: I2 (D8)
2606.13232	tactile	0	0	25	included
2606.14218	torque	4	2	20	included
2606.17055	tactile	–	–	–	excluded: I2 (D8)
2606.24450	mixed	–	–	–	excluded: I2
2606.30988	force	1	0	15	included
2607.02503	tactile	6	0	20	included
2607.03387	tactile	10	0	25	included
2607.03529	current	14	6	52	included
2607.03723	tactile	8	8	20	included
2607.07287	tactile	12	1	100	included
2607.08575	none	–	–	–	excluded: I1
2607.11734	current	2	0	40	included
2607.14236	force	1	1	30	included
2607.14578	current	6	2	20	included
2607.18231	force	6	2	18	included
2607.20683	tactile	8	0	20	included
2607.20912	force	2	2	20	included
2607.22530	tactile	8	4	50	included
2608.01102	tactile	0	0	20	included
2608.01733	tactile	–	–	–	excluded: I2
2608.01824	tactile	1	0	20	included
2608.07558	mixed	–	–	–	excluded: survey
2608.15009	force	10	1	45	included
2608.15560	tactile	4	0	11	included
2608.24603	none	–	–	–	excluded: I1
2608.25798	tactile	6	0	16	included
2608.25872	tactile	12	0	10	included
2609.02020	torque	–	–	–	excluded: I2
2609.03889	torque	6	6	25	included
2609.05832	force	0	0	25	included
2609.07002	tactile	5	2	200	included
2609.07747	tactile	1	1	30	included
2609.12549	tactile	6	1	20	included
2609.13779	torque	–	–	–	excluded: I2
2609.14156	tactile	8	2	30	included
2609.15012	force	–	–	–	excluded: I2 (D8)

arXiv ID	Signal	Binary	Resolved	Median n	Status
2609.15726	tactile	–	–	–	excluded: I2 (sim.)
2609.15921	tactile	4	4	30	included
2609.16437	tactile	–	–	–	excluded: I2
2609.16504	tactile	11	0	10	included
2609.16586	none	–	–	–	excluded: I1
2609.18164	torque	–	–	–	excluded: I2 (sim.)
2609.18165	tactile	–	–	–	excluded: hardware only
2609.18242	torque	12	1	30	included
2609.18497	force	12	5	30	included
2609.19137	force	8	2	10	included
2609.20649	tactile	–	–	–	excluded: I2 (D8)
2609.20980	tactile	12	5	80	included
2609.21369	none	–	–	–	excluded: I1
2609.22840	force	3	3	30	included
2609.22895	none	–	–	–	excluded: I1
2609.23614	force	–	–	–	excluded: I1
2609.23800	mixed	3	3	50	included
2609.23888	current	3	2	20	included
2609.23968	force	–	–	–	excluded: I2
2609.24976	tactile	4	2	20	included
2609.25785	force	3	3	20	included
2609.27381	current	–	–	–	excluded: I2
2609.29017	current	0	0	30	included
2609.29822	tactile	8	4	60	included
2609.30969	tactile	1	0	20	included
2609.31924	force	3	3	20	included
2609.33882	tactile	–	–	–	excluded: I1
2609.34006	tactile	1	1	20	included
2609.34182	tactile	–	–	–	excluded: I2 (D8)
2609.34286	tactile	–	–	–	excluded: I2
2609.34400	force	–	–	–	excluded: I2
2609.36596	tactile+torque	10	7	20	included
2609.36785	tactile	–	–	–	excluded: I1 (D6)
2609.37433	force	8	2	25	included
2609.37552	force	5	3	50	included
2609.39017	tactile	–	–	–	excluded: I3
2610.00638	tactile	10	0	20	included