BRUSHWORKS · KILL-LOG SUMMARY · LAUNCH EDITION · SEPTEMBER 2026

Kill-log summary.

The models we did not ship, and why

Every model retired at the publish gate or in the field, in aggregate, by family and by reason, with the baselines that beat them, what shipped and how it performed, the patterns a year reveals, and the cards of a retired and a shipped model.

Brushworks

brwrks.com/reports
research@brwrks.com
Cite as: Brushworks (2026), *Kill-log summary*, Launch edition · September 2026.

Contents

Section 1 Why the firm publishes what it rejected

Section 2 The gate, in brief

Section 3 Ledger by family

Section 4 Ledger by reason

Section 5 What shipped, and how it performed

Section 6 Patterns and lessons

Section 7 What changes next year

Appendix A A retired model's card

Appendix B A shipped model's card

Section 1

Why the firm publishes what it rejected

A model catalog is only as honest as the list of what it rejected. Every data firm shows a buyer the models that worked; almost none will show the ones that did not, and the omission is not small. A vendor that trains a hundred models and reports the twelve that worked is describing a selection, not a method. This summary publishes, in aggregate, the models Brushworks retired at the publish gate or in the field, the reasons, the baselines that beat them, and what the pattern says about the work.

LAUNCH EDITION

The firm publishes this summary each year with the annual report. This edition establishes the format and the definitions, and populates the ledgers with representative proportions drawn from the firm's programs so that the shape of a year is visible. The first audited count — for the year ending September 30, 2026 — is published in spring 2027. Read the numbers here as the method's proportions, not as audited totals.

1.1 What a buyer should ask

Not "what is your win rate" but: how many candidate models did you test this year; how many beat a competently used voter file; how many beat the platforms' own look-alike targeting; how many did you retire for reconstructing an attribute you say you do not use; how many shipped; how many did you kill in the field, and when. A ratio of tested to shipped near one is a catalog without a gate. The firm's is under one in two.

Section 2

The gate, in brief

The full standard is in the methodology notes, Section 8. A candidate model must, on the client's own sealed holdout: beat the voter file used well by a pre-registered margin (0.03 in AUC and ten percent relative in top-decile lift by default); beat the delivery platforms' look-alike targeting seeded with the campaign's supporters by the same margin; and fail to reconstruct any excluded attribute — ethnicity, religion, ideology — by proxy, measured as a classifier trained on the model's outputs achieving no more than 0.60 AUC on a labeled reference set. A model that passes is shipped with a card and a field kill threshold; a model whose seven-day rolling discrimination falls below that threshold for three consecutive days is retired in field and the cohort is rescored.

1
Train
On recorded outcomes from prior programs, time-split

2
Test on the holdout
The client's own sealed control

3
Beat the file
Participation, party, precinct, age, prior contact

4
Beat look-alike
Platform expansion seeded with supporters

5
Pass the proxy test
No excluded attribute reconstructed

6
Ship with a card
Watched daily; killed on drift

Section 3

Ledger by family

FAMILY	TESTED	FAILED THE FILE	FAILED LOOK-ALIKE	PROXY RECONSTRUCTION	SHIPPED	KILLED IN FIELD	SURVIVED THE YEAR
Behavioral enrichment	22	8	3	1	10	2	8
Dispositions	19	5	2	1	11	3	8
Occupational	14	7	1	0	6	1	5
Interests	16	9	2	1	4	0	4
Affinities	13	6	1	0	3	1	2
All families	**84**	**35**	**9**	**3**	**34**	**7**	**27**

Table 3.1 Candidate models by family and fate, representative year. Failed-the-file models are retired at the first baseline; failed-look-alike at the second; proxy reconstruction retires a model regardless of its position. Thirty-four shipped; seven were killed in field; twenty-seven ended the year in use.

3.1 Reading the ledger

Two-fifths of candidates fail the first baseline: a competently used voter file is a strong predictor of participation and a fair predictor of persuadability, and a model that cannot beat it has nothing to add. Interests models fail most often, because interest signals predict attention rather than acts. Dispositions models ship most often and are killed in field most often, because dispositions move with events and a model trained before a shock can be wrong after it. Occupational models are the most stable once shipped.

Section 4

Ledger by reason

REASON	COUNT	SHARE OF TESTED	DEFINITION	REPRESENTATIVE CASE (ANONYMIZED)
Failed the file	35	42%	Did not beat the voter file used well on the client's holdout for the pre-registered outcome	E-14, an enrichment model of turnout propensity: AUC 0.69 against the file's 0.70. The file was used instead; the report said so

REASON	COUNT	SHARE OF TESTED	DEFINITION	REPRESENTATIVE CASE (ANONYMIZED)
Failed look-alike	9	11%	Beat the file but not the platforms' expansion audience for the outcome	A-05, an affinities model for contribution: AUC 0.66 against look-alike 0.68 on the client's supporter seed; not used for that channel
Proxy reconstruction	3	4%	A classifier on the model's outputs recovered an excluded attribute above 0.60 AUC	I-09, an interests model whose outputs reconstructed religious affiliation at 0.64 through grocery and calendar signals; retired at the gate despite passing both baselines
Killed in field	7	8% (21% of shipped)	Seven-day rolling AUC below the card's threshold for three consecutive days	D-07, a dispositions model of economic anxiety, fell from 0.71 to 0.63 in the week after a plant closure dominated local news; the cohort was rescored by the file at 0.66
Shipped and survived	27	32%	Passed the gate; held above threshold for the program's life	D-02, a dispositions model of competence-under-pressure responsiveness: AUC 0.74 against the file's 0.68 and look-alike's 0.70, stable across three statewide programs

4.1 Field kills, by cause

CAUSE	KILLS	WHAT HAPPENED	WHAT CHANGED
Event shock	3	A local or national event moved the disposition the model predicted	Dispositions models now carry an event-sensitivity note on the card and a shorter three-day rule
Population drift	2	Late registrants and movers changed the cohort's composition mid-program	Cohorts are re-sized at week four and new members receive a fresh holdout draw
Signal outage	1	A source paused refresh for eleven days; the model's inputs went stale	Provenance records now carry a refresh-cadence alarm
Seasonality	1	An affinities model trained on autumn behavior underperformed in a spring special	Training windows are matched to the program's season

Section 5

What shipped, and how it performed

Shipped models, at the gate, discriminated the primary outcome at AUC between 0.71 and 0.78 against file baselines of 0.66 to 0.71 and look-alike baselines of 0.67 to 0.72. Top-decile lift over the file ranged from twelve to thirty-one percent. In field, surviving models held their gate discrimination within 0.03 for the life of their programs; the seven kills are the cases that did not. The numbers are the reason the firm reports lift against a holdout rather than model accuracy: a model that is three points better than the file is worth a great deal at the scale of a state and nothing to a buyer who cannot see it.

5.1 Baseline discrimination, for reference

BASELINE	OUTCOME	AUC RANGE	NOTE
The file used well	Turnout	0.72–0.79	Participation history is a strong predictor; hard to beat for mobilization
The file used well	Persuadability (panel)	0.58–0.64	Weak; where behavioral models add most
Look-alike	Contribution	0.66–0.72	Strong on supporters' near neighbors; weak beyond them
Look-alike	Persuadability	0.57–0.62	Expansion audiences find the similar, not the persuadable

Section 6

Patterns and lessons

1. **The file is the model to beat for turnout, and usually wins.** Most enrichment candidates for mobilization fail it; the firm's value in mobilization is timing and register, not scoring.
2. **Behavior beats look-alike for persuasion by a wide margin, and look-alike beats behavior for finding donors' twins.** The catalog is used channel by channel, not as a single ranking.
3. **Proxy reconstruction is rare but never zero.** Three models in a representative year passed both baselines and would have reconstructed religion or ethnicity. The test exists because performance and prohibition are independent.
4. **Dispositions models are the most valuable and the most fragile.** They ship most and die most. The daily watch is the price of using them.
5. **Most field kills are events, not errors.** A model that was right on Monday and wrong on Friday was not a bad model; it was a model whose world changed. The kill log records the world.

Section 7

What changes next year

- Audited counts, reconciled to program kill logs by the independent reviewer, replace representative proportions.
- The ledger adds a column for models retired before the gate for data-quality reasons, which this edition folds into "failed the file".
- Field kills are reported with the days between threshold breach and retirement, so a buyer can see the three-day rule in operation.
- Program-level kill logs continue to be delivered to each client and held under agreement; nothing in this summary identifies a program.

Appendix A

A retired model's card

FIELD	CONTENT
Identifier and family	I-09, interests
Outcome predicted	Response to a cost-of-living frame, panel favorability, statewide program
Training basis	Four programs, 2022–2024; time-split at January 2025
Validation	AUC 0.72 (file 0.61; look-alike 0.60); top-decile lift 24% over the file — passed both baselines
Proxy test	Religious affiliation reconstructed at AUC 0.64 (threshold 0.60) through grocery-category and weekly-calendar signals — **failed**
Decision	Retired at the gate, March 2026; the signals were reviewed and two were removed from the registry's candidate set for disposition models
Owner	Head of research

Appendix B

A shipped model's card

FIELD	CONTENT
Identifier and family	D-02, dispositions
Outcome predicted	Responsiveness to a competence-under-pressure frame; panel favorability
Training basis	Six programs, 2021–2025; time-split at October 2025
Validation	AUC 0.74 (file 0.68; look-alike 0.70); top-decile lift 19% over the stronger baseline; margin 0.04 — passed
Proxy test	Ethnicity 0.53; religion 0.51; ideology 0.55 — passed
Permitted use	The Night Nurse and the Steward cohorts; all registers; United States
Known limits	Weaker discrimination among voters under thirty; not validated for institutional outcomes
Field kill threshold	Seven-day AUC below 0.70 for three consecutive days
Owner and date	Head of research; card issued November 2025; in use

Kill-log summary, launch edition, September 2026. Next edition: with the annual report, spring 2027, with audited counts.