Would you take a +923 point profit if it meant sitting through a -1,129 drawdown, or +1,718 with only -310?

Inside our proposal: how SAIL Lab's own AI balances profit and risk

XAUUSD Trend Policy over 14 months (Jan 2025 - Feb 2026): training window, re-optimization frequency, entry filters and ML

A strategy is only usable if you can live through its worst stretch. We start from the Baseline Trend Policy (config.py with fixed parameters) and add SAIL Lab's proposal step by step: walk-forward Optimization, Entry filters, and our ML system, then measure what each step does to profit and to drawdown. Optimization: every R trading days the parameters are re-chosen from 3,001 sets using only the previous W days, then traded until the next re-optimization; every start offset (phase) is simulated separately, tick by tick with live_trader.py rules. This repeats our 7-month study (Mar-Oct 2026) in the opposite market: here gold rose 101% (2,625 to 5,279), there it fell 21%.

Source: 95.8M Exness XAUUSD ticks, 2025-01-01 to 2026-02-27 (302 trading days, only Good Friday 2025-04-18 missing), downloaded from MT5 with Data/download_tick.py. None of these days were used in the 7-month study (Mar-Oct 2026). Live-trader rules, points = price units (1 pt = $1 per 0.01 lot, so $4 at the committed config.py's LOT_SIZE 0.04).

+923 pts / -1,129
Baseline Trend Policy: config.py held fixed (32th percentile of 3,001 fixed sets). Profit / max drawdown: 0.8

+2,610 pts / -702
Baseline + Our Optimization. Profit / max drawdown: 3.7

+2,997 pts / -774
Baseline + Our Optimization + Our Entry filters. Profit / max drawdown: 3.9

+1,718 pts / -310
Baseline + Our Optimization + Our Entry filters + Our ML: the most profit per point of risk. Profit / max drawdown: 5.5

+2,163 pts / -665
Gold buy-and-hold, 3,021 to 5,279 (+75%). Profit / max drawdown: 3.3

Each card: out-of-sample net P&L / max drawdown, points. Re-optimization every 5 days on the last 60 days, averaged over phases. From the Baseline to the full proposal, the drawdown falls from 1129 to 310 (-73%) while the profit still nearly doubles. Gold's drawdown uses every 2nd day.

Main result: the 60-day re-optimization and our ML hold in both a falling and a rising market

A 60-day training window scored on live-rule P&L or plateau was the best schedule in both markets (here it beats 93-98% of fixed sets at every interval), and short windows were the worst relative to the field. Our ML lowered the drawdown in both. How often you re-optimize made no difference here. The entry filters, the "longs are weak" finding and the trend-gate verdict depend on the market: they worked while gold fell and failed or reversed while it rose. Absolute numbers are higher here only because almost every parameter set made money while gold doubled.

Which findings hold in both a falling and a rising market?

Finding while gold fell 21% (Mar-Oct 2026)	While gold rose 101% (Jan 2025 - Feb 2026)	Verdict
40-60 day training windows beat 5-20 day windows	60-day windows beat 80% of fixed sets on average, 5-10 day windows 51%. 20-day windows now also do well (77%).	Held
Re-optimize every 20-60 days, not daily or every 5 days	No difference: 66-70% of fixed sets beaten at every interval.	Not confirmed
Score on live-rule P&L, not the current pipeline score	Pipeline score beat 52% of fixed sets on average (coin flip), live P&L 75%, plateau 76%.	Held
Skip 14-17 UTC + only trade with today's move improves every set	Picked again on the first half (New York-time version), it improved only 32% of sets on the second half; median set -70 pts.	Failed
The long side is weak	Committed config, test period: longs +1,202, shorts -278. In a rising market the longs carried the strategy.	Reversed
Each side only with its 20-day trend is harmful	The 10-day version improved 100% of sets on the second half, but cut the first half by 329 pts (median).	Reversed
Daily loss stops and max losses per day hurt	Stop after 4 losses/day improved 81% of sets on the second half but cost 261 pts (median) on the first half.	Mixed
Our ML trades some profit for much lower risk	On top of Optimization + Entry filters: profit -1,279 pts (every 5 days) and -884 (every 20 days), max drawdown 774 to 310 and 719 to 322, well below random trade dropping at the same profit.	Held in both
In-sample rank barely predicts the next 20 days	Live-P&L IC -0.04 to +0.09; the pipeline score is negative (-0.08 to -0.11).	Held

Same scripts and settings in both markets, no overlapping days. "Held": true in both markets; "Failed" / "Reversed": true only while gold fell. "Fixed sets beaten" is the share of the 3,001 fixed parameter sets whose out-of-sample total a schedule beat.

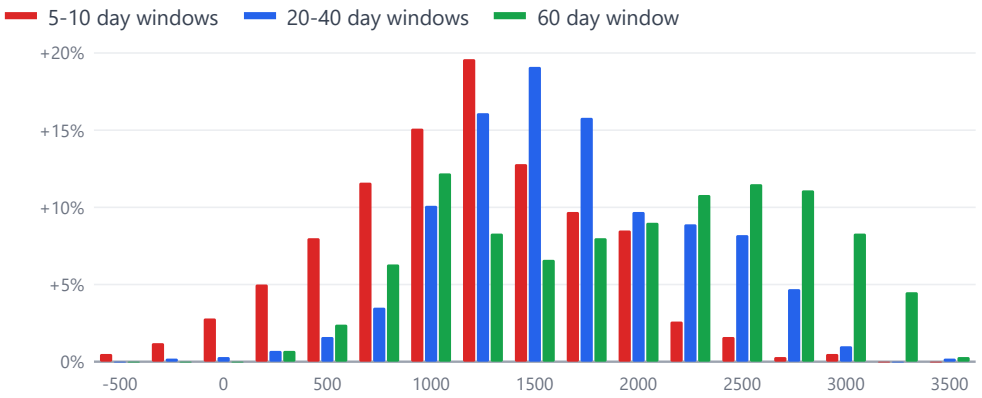
1. Impact of training window and frequency

Out-of-sample P&L of 3,001 fixed parameter sets, with re-optimization schedules marked



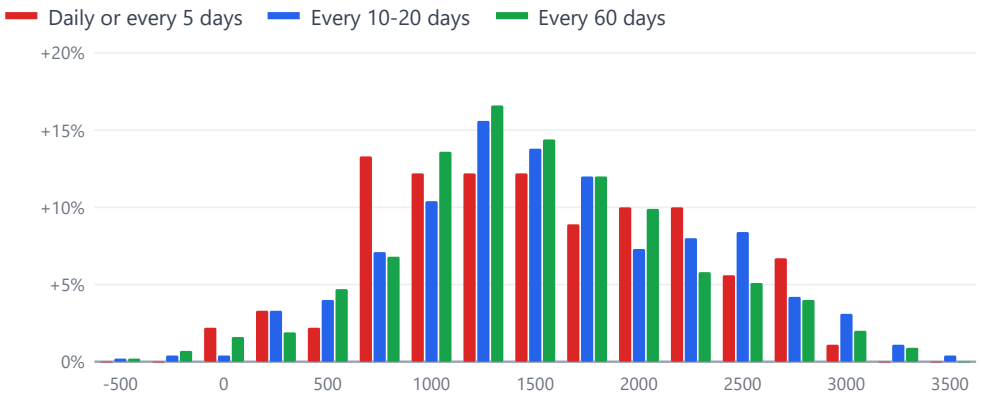
Bars: number of fixed parameter sets per 119-pt bin. Lines: walk-forward schedules (average over phases) and the committed config. Green = best schedules, red = worst schedule and the committed config, blue = pipeline score, dashed = median. Fixed sets: 10th percentile +518, 90th +2,038, best +4,146. The working-copy config.py (25/85/13) made +1,016 (38th percentile). Out-of-sample test: 2025-03-26 to 2026-02-27 (242 trading days); the first 60 days are training-only.

Outcome distribution by training window



x: out-of-sample net P&L, pts (left edge of a 250-pt bin). y: % of walk-forward outcomes (3 selection rules x 5 intervals x every phase). Mean: 5 d +1,287, 10 d +1,332, 20 d +1,689, 40 d +1,902, 60 d +2,074. Share above the fixed-set median: 59% / 61% / 84% / 92% / 82%.

Outcome distribution by re-optimization interval



x: out-of-sample net P&L, pts. y: % of walk-forward outcomes. Mean: daily +1,647, 5 d +1,639, 10 d +1,687, 20 d +1,741, 60 d +1,625. The distributions overlap almost completely: in this period frequency did not matter.

Window and frequency together

Group	Outcomes	Mean (pts)	Median	Above fixed median	10th pct	90th pct
Short window (5-10 days), re-opt every 1-5 days	36	+1,092	+1,040	42%	+382	+1,768
Short window (5-10 days), re-opt every 20-60 days	480	+1,340	+1,365	63%	+571	+2,073
Long window (40-60 days), re-opt every 1-5 days	36	+2,030	+2,204	89%	+1,081	+2,806
Long window (40-60 days), re-opt every 20-60 days	480	+1,966	+1,887	86%	+1,077	+2,930

Outcomes = rule x window x interval x phase combinations. The fixed-set median over the same period is +1,176 pts. Window length splits the groups; interval does not.

Everything made money here, so look at rank, not sign. Short windows (5-10 days) beat only about half of the fixed sets whatever the interval; 20-60 day windows beat about 78%.

Your old daily_auto_trade.bat style (5-day window re-chosen daily by the pipeline score) made +1,549, 72th percentile. Its neighbors (every 5 days: +1,118; 10-day window daily: +1,514) show again that this is noise.

The pipeline score (calculate_score on tick_engine) is the weak link in both periods: with a 60-day window it beat only 50% of fixed sets, while live-rule P&L and plateau beat 95% and 96%.

Score = net P&L with live rules

Training window	Daily	Every 5 d	Every 10 d	Every 20 d	Every 60 d
5 days	+2,035 (90%)	+1,289 (54%)	+1,325 (57%)	+1,425 (64%)	+1,580 (69%)
10 days	+838 (26%)	+1,370 (57%)	+1,530 (64%)	+1,483 (63%)	+1,618 (70%)
20 days	+1,374 (62%)	+1,904 (86%)	+1,718 (78%)	+1,734 (72%)	+1,494 (65%)
40 days	+1,902 (86%)	+1,991 (87%)	+2,137 (88%)	+2,044 (86%)	+1,884 (79%)
60 days	+2,486 (96%)	+2,610 (97%)	+2,482 (96%)	+2,448 (93%)	+2,421 (94%)

Score = current pipeline (calculate_score on tick_engine)

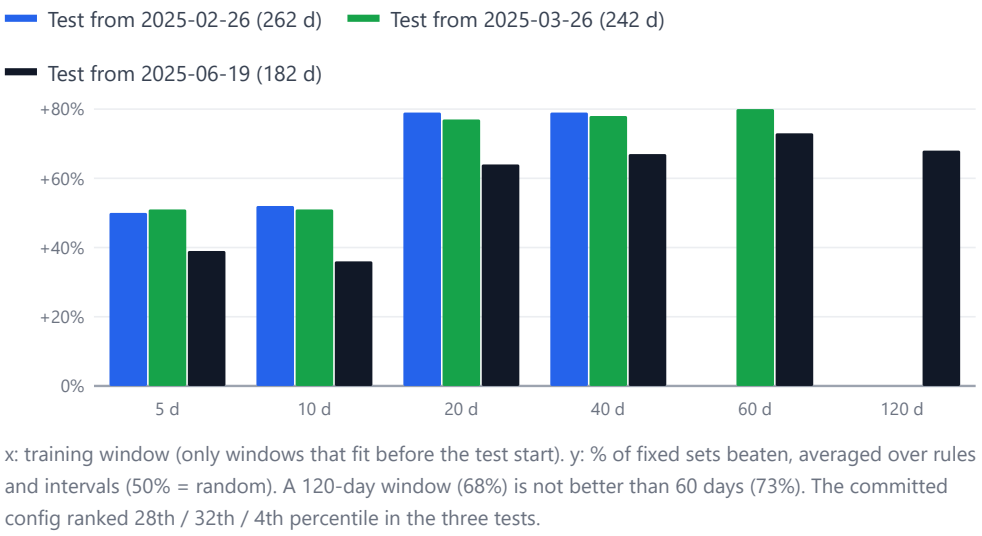
Training window	Daily	Every 5 d	Every 10 d	Every 20 d	Every 60 d
5 days	+1,549 (72%)	+1,118 (44%)	+958 (36%)	+984 (36%)	+870 (34%)
10 days	+1,514 (70%)	+722 (26%)	+791 (26%)	+841 (31%)	+872 (34%)
20 days	+1,614 (76%)	+1,672 (76%)	+1,490 (68%)	+1,662 (76%)	+1,346 (58%)
40 days	+500 (9%)	+1,378 (62%)	+1,531 (70%)	+1,678 (75%)	+1,434 (66%)
60 days	+1,434 (66%)	+1,090 (43%)	+1,133 (46%)	+1,153 (47%)	+1,156 (48%)

Score = plateau (30 nearest sets averaged)

Training window	Daily	Every 5 d	Every 10 d	Every 20 d	Every 60 d
5 days	+850 (27%)	+636 (21%)	+941 (39%)	+1,401 (61%)	+1,156 (48%)
10 days	+1,040 (40%)	+1,160 (49%)	+1,601 (66%)	+1,622 (69%)	+1,601 (66%)
20 days	+2,484 (96%)	+2,379 (94%)	+2,121 (88%)	+2,077 (87%)	+1,941 (80%)
40 days	+2,413 (95%)	+2,362 (95%)	+2,713 (97%)	+2,722 (97%)	+2,669 (97%)
60 days	+2,669 (97%)	+2,904 (98%)	+2,833 (97%)	+2,844 (97%)	+2,930 (97%)

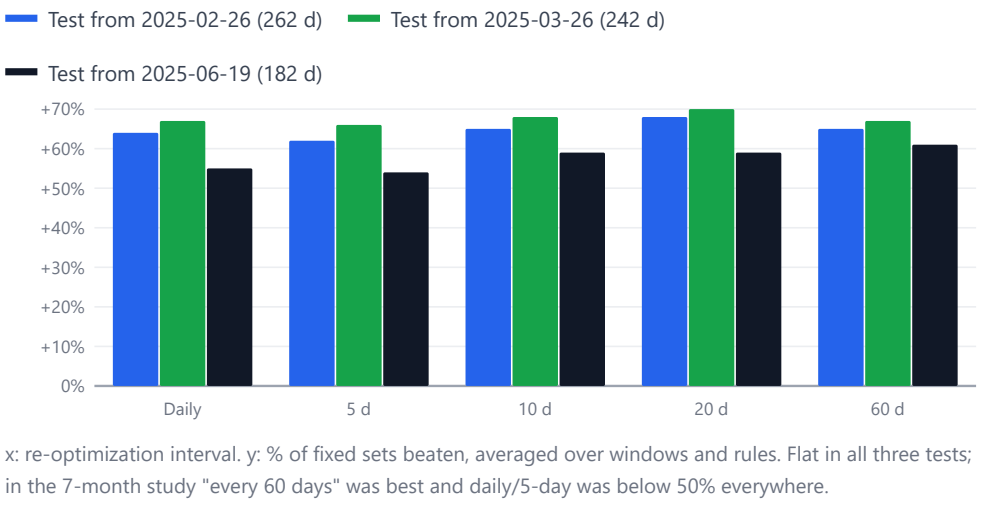
Cells: out-of-sample net pts averaged over phases (and % of the 3,001 fixed sets that result beat). Rows: training window. Columns: re-optimization interval. The 60-day row is the best row for live P&L and plateau; for the pipeline score no row is reliably good.

Is the window effect stable when the test period moves?



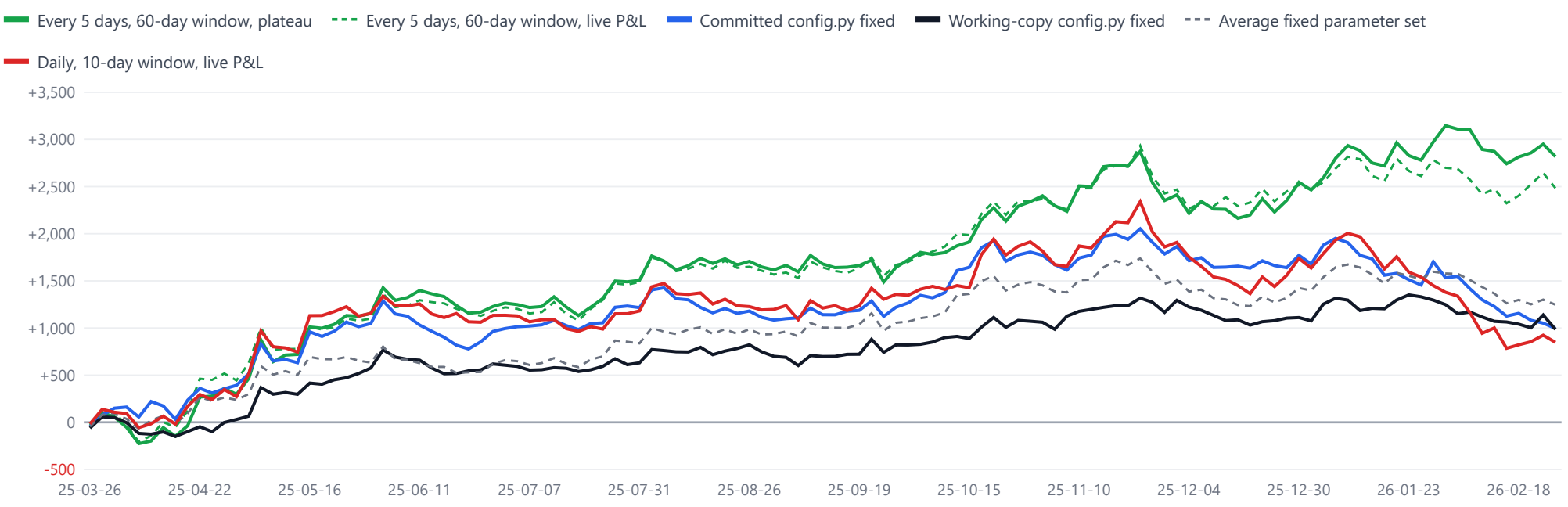
x: training window (only windows that fit before the test start). y: % of fixed sets beaten, averaged over rules and intervals (50% = random). A 120-day window (68%) is not better than 60 days (73%). The committed config ranked 28th / 32th / 4th percentile in the three tests.

Is the frequency effect stable when the test period moves?



x: re-optimization interval. y: % of fixed sets beaten, averaged over windows and rules. Flat in all three tests; in the 7-month study "every 60 days" was best and daily/5-day was below 50% everywhere.

Cumulative out-of-sample P&L



x: trading day (YY-MM-DD, UTC, every 2nd day). y: cumulative net P&L, points. Big gains in May and October 2025; February 2026 (gold +478 in a very volatile month) was the worst month for almost everything.

Monthly out-of-sample P&L (points)

Schedule	Mar 25 (26-31)	Apr 25	May 25	Jun 25	Jul 25	Aug 25	Sep 25	Oct 25	Nov 25	Dec 25	Jan 26	Feb 26	Total	Months +
Every 5 days, 60-day window, plateau	+106	+192	+856	+76	+275	+97	+56	+744	-50	+112	+532	-92	+2,904	10 / 12
Every 5 days, 60-day window, live P&L	+146	+298	+655	+86	+304	+77	+106	+697	+58	+40	+289	-147	+2,609	11 / 12
Every 20 days, 60-day window, live P&L	+146	+171	+772	+67	+287	+100	+143	+570	-72	+58	+344	-139	+2,447	10 / 12
Every 20 days, 40-day window, live P&L	+146	+34	+737	+103	+229	+42	+93	+523	-18	+96	+384	-324	+2,045	10 / 12
Daily, 5-day window, pipeline score (like daily_auto_trade.bat)	+100	+290	+462	-134	+37	+116	+224	+229	-48	+22	+342	-92	+1,548	9 / 12
Daily, 60-day window, pipeline score	+92	+338	+420	-63	+237	+93	-153	+171	-33	0	+546	-215	+1,433	7 / 12
Every 5 days, 5-day window, plateau	+11	+233	+466	-37	+128	+53	+135	+273	-138	-62	+19	-443	+638	8 / 12
Daily, 10-day window, live P&L	+157	+115	+884	-28	+52	-4	+172	+467	+43	-223	-240	-559	+836	7 / 12
Committed config.py, fixed (45/60/5)	+110	+283	+655	-131	+299	-135	+184	+506	+14	-107	-38	-717	+923	7 / 12
Working-copy config.py, fixed (25/85/13, uncommitted)	+75	-45	+546	+15	+40	+62	+102	+265	+106	-91	+147	-207	+1,015	9 / 12
Average of 3,001 fixed parameter sets	+92	+146	+395	-16	+220	+68	+164	+384	+15	-71	+167	-316	+1,248	9 / 12

May and October 2025 alone are +779 pts for the average set. The committed config lost -717 in February 2026, more than the average set (-316).

Does in-sample rank predict the future?



x: training window. y: mean Spearman correlation between in-sample score and future P&L across 3,001 sets. Best: 60-day window, next 60 days, 0.146 (t 1.86). The pipeline score is negative at every window.

78% of a parameter set's daily P&L variance is the market (the average of all sets that day), not the parameters (69% in the 7-month study). Rank correlations between the four ~75-day quarters range from -0.098 to 0.537.

Daily P&L has no persistence (autocorrelation -0.057), weekly -0.003, monthly -0.044. As before, a good week says nothing about the next one, which is why 5-10 day windows do no better than picking a set at random.

A 60-day window helps because it averages over several market states and picks parameters that are robust. Of everything tested, this is the one result that now holds in a falling (2026) and a rising (2025) market.

2. Where the strategy fails

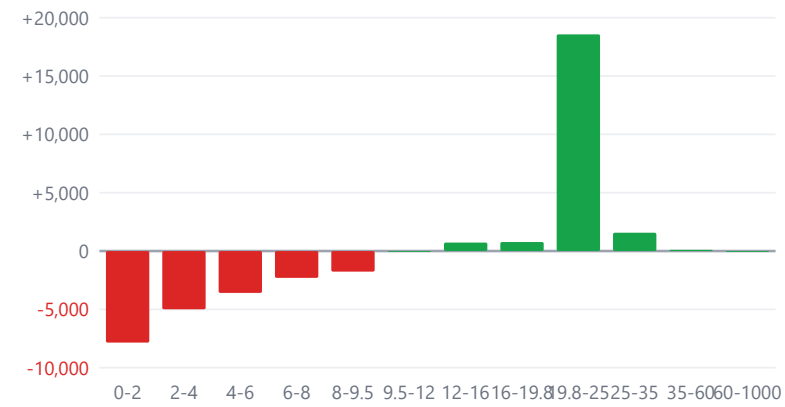
Committed config.py over all 302 days with live rules: 4,619 trades, +1,262 pts, PF 1.06, win rate 31%, max drawdown 1192 pts. Under tick_engine.py rules (what the optimizer uses) the same parameters make -1,408 pts, and 1,013 trades are dropped as open at day end.

This time the long side made the money

Month	Gold open to close	Move (pts)	Long trades	Long net	Short trades	Short net
Jan 25	2,625 to 2,799	+174	119	+441	79	-150
Feb 25	2,803 to 2,858	+56	123	+178	129	+188
Mar 25	2,874 to 3,121	+247	160	+218	95	-428
Apr 25	3,123 to 3,273	+150	217	0	154	+283
May 25	3,275 to 3,290	+15	191	+517	198	+139
Jun 25	3,308 to 3,312	+4	151	-235	150	+104
Jul 25	3,312 to 3,289	-23	121	+13	132	+286
Aug 25	3,288 to 3,448	+160	122	+90	105	-226
Sep 25	3,445 to 3,863	+418	199	+400	108	-216
Oct 25	3,864 to 4,002	+137	277	+467	176	+39
Nov 25	3,996 to 4,223	+227	213	+227	167	-213
Dec 25	4,222 to 4,319	+97	210	-102	174	-4
Jan 26	4,334 to 4,892	+558	272	+56	145	-94
Feb 26	4,801 to 5,279	+478	248	-400	184	-318

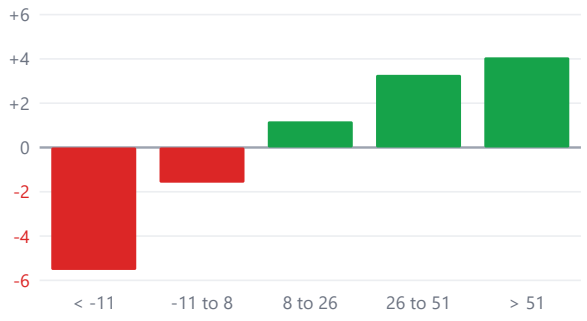
Longs: 2,623 trades, +1,870 pts, positive in 10 of 14 months. Shorts: 1,996 trades, -610 pts. In Mar-Oct 2026 it was the other way round (longs -1,025, shorts +1,880). The side that wins is the side of the trend, so neither side is broken.

Exit anatomy: net P&L by maximum favorable excursion



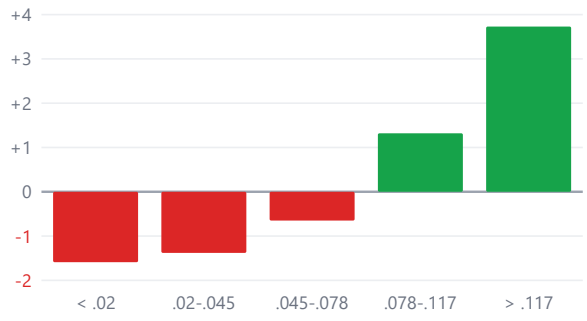
x: furthest the trade went in your favor, pts. y: total net P&L of trades in that bucket. 834 trades never got 2 pts in favor (-7,846 pts). Winners are capped near 20 pts by the 1-pt trailing distance (median capture 94% of MFE); only 65 trades ran past 25 pts. Same shape as in 2026.

Trades against the day's direction lose (hindsight)



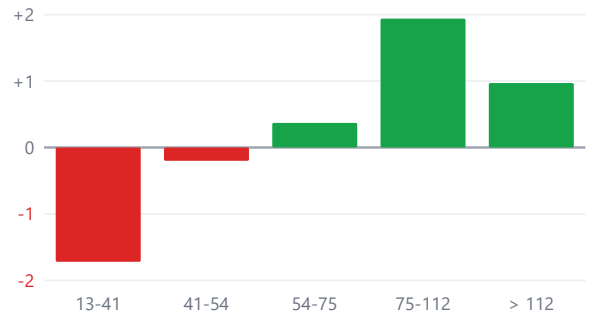
x: day's net move in trade direction, pts. y: mean P&L per trade, pts (~924 trades per bucket, std. error 0.32-0.46 pts). Same as in 2026: the day type decides.

Choppy days lose (hindsight)



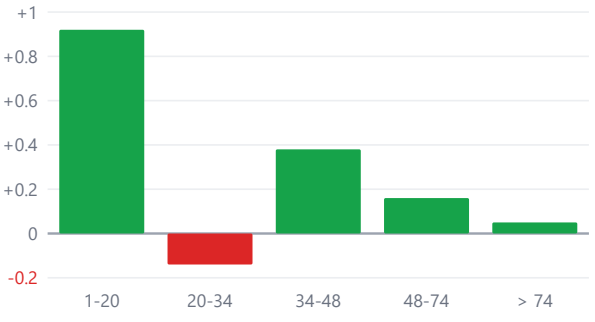
x: day efficiency ratio |close-open| / path. y: mean P&L per trade, pts (~924 trades per bucket, std. error 0.37-0.45 pts). Same as in 2026.

Quiet days lose (hindsight)



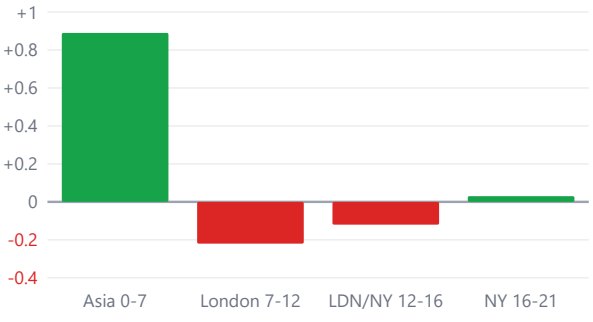
x: day high-low range, pts. y: mean P&L per trade, pts (~924 trades per bucket, std. error 0.32-0.46 pts). Same as in 2026, at lower price levels.

Late entries: no longer a clear effect (causal)



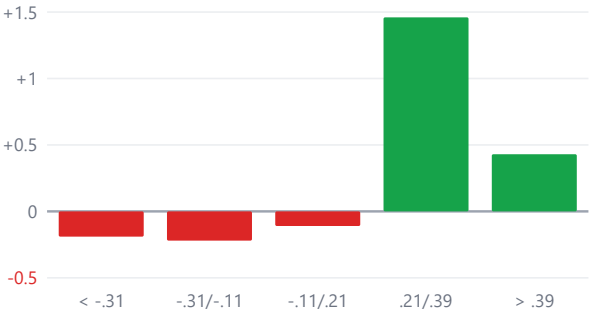
x: day's range already traveled at entry, pts. y: mean P&L per trade, pts (~924 trades per bucket, std. error 0.38-0.44 pts). In 2026 entries after ~38 pts lost; here only the first bucket stands out.

Session: the overlap effect is gone (causal)



x: UTC session at entry. y: mean P&L per trade, pts (~1141 trades per bucket, std. error 0.32-0.41 pts). 12-16 UTC lost -0.94 pts/trade in 2026; -0.12 here.

Pull-distance band: flipped sign (causal)



x: close minus pull EMA at the signal, ATR units. y: mean P&L per trade, pts (~924 trades per bucket, std. error 0.39-0.42 pts). The 0.13-0.31 ATR band lost in 2026; the 0.21-0.39 band is the best here. Noise.

The three hindsight patterns (direction, chop, quiet days) repeat in both periods; they describe the problem but cannot be traded. The causal patterns found in 2026 (late entries, the London/New York overlap, the pull-distance band) mostly did not repeat, which is consistent with section 5.

---

3. ML filter and simple day filters

ML trade filter by training window and retraining interval

Model	Training window	Daily	Every 5 d	Every 20 d	Every 60 d
Logistic regression	10 days	-1,837 (0.508)	-1,350 (0.518)	-372 (0.533)	-873 (0.526)
Logistic regression	20 days	-1,059 (0.537)	-475 (0.546)	+332 (0.544)	-857 (0.515)
Logistic regression	40 days	-144 (0.555)	-16 (0.565)	+12 (0.556)	-123 (0.531)
Logistic regression	60 days	-304 (0.556)	-114 (0.560)	-308 (0.552)	-758 (0.537)
Logistic regression	All history	-154 (0.557)	+260 (0.564)	+9 (0.564)	+34 (0.556)
Gradient boosting	10 days	-1,888 (0.493)	-1,493 (0.505)	-392 (0.533)	-64 (0.515)
Gradient boosting	20 days	-846 (0.520)	-476 (0.526)	-298 (0.526)	-744 (0.511)
Gradient boosting	40 days	-233 (0.538)	+64 (0.551)	+53 (0.541)	-660 (0.519)
Gradient boosting	60 days	-28 (0.551)	+95 (0.552)	+171 (0.555)	-349 (0.534)
Gradient boosting	All history	-192 (0.544)	-53 (0.547)	+433 (0.550)	-153 (0.536)

Committed config's 3,950 test trades, 140 features. Cells: P&L of the trades kept minus P&L of all trades, pts (and out-of-sample AUC). Take a trade if P(win) >= the training win rate. Only 10 of 40 settings improved on taking every trade.

Mean AUC 0.544 (logistic regression) and 0.532 (gradient boosting), close to 2026's 0.523. The model ranks trades slightly better than chance, but the trades it skips were net profitable, so skipping them costs P&L.

Short windows (10 days) are again the worst, by far: up to -1,888 pts. 40-day, 60-day and all-history windows lose less (-758 to +433). This single model on a fixed parameter set cut P&L in both markets; section 5 tests our ML system retrained inside the re-optimization, where it lowers risk.

Day-level direction filters, decided before the day starts

Filter	Trades	Net (pts)	Random filter, same size	Long net	Short net
no filter	3957	+923	+923	+1,202	-278
side momentum: trade a side only if its last 3d P&L > 0	1931	-68	+451	+3	-71
side momentum: trade a side only if its last 5d P&L > 0	2082	-206	+486	+271	-476
side momentum: trade a side only if its last 10d P&L > 0	2261	+205	+528	+619	-414
side momentum: trade a side only if its last 20d P&L > 0	2281	+611	+532	+763	-152
daily trend: long only if yesterday close > close 3d earlier, else short only	2102	+619	+491	+546	+74
daily trend: long only if yesterday close > close 5d earlier, else short only	2079	+550	+485	+477	+73
daily trend: long only if yesterday close > close 10d earlier, else short only	2036	+461	+475	+719	-258
daily trend: long only if yesterday close > close 20d earlier, else short only	2151	+46	+502	+319	-274
daily trend: long only if yesterday close > close 60d earlier, else short only	2243	+1,061	+523	+1,101	-40

Committed config, Mar 26 2025 - Feb 27 2026. The side-momentum filters that beat a random filter by +270 to +340 pts in 2026 lose to it here (3-5 day versions: -520 to -690 pts vs random). The 60-day daily trend filter is the best (+1,061 vs +523 random): long only while gold is above its level of 60 days ago.

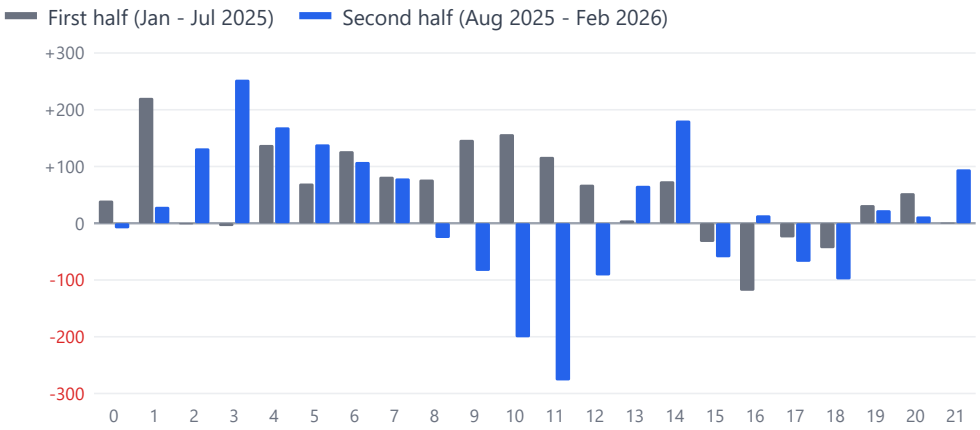
The config.py history section of the 7-month canvas is not repeated: the committed versions in git start in May 2026, after this period.

---

4. Limiting losses: entry filters tested out-of-sample

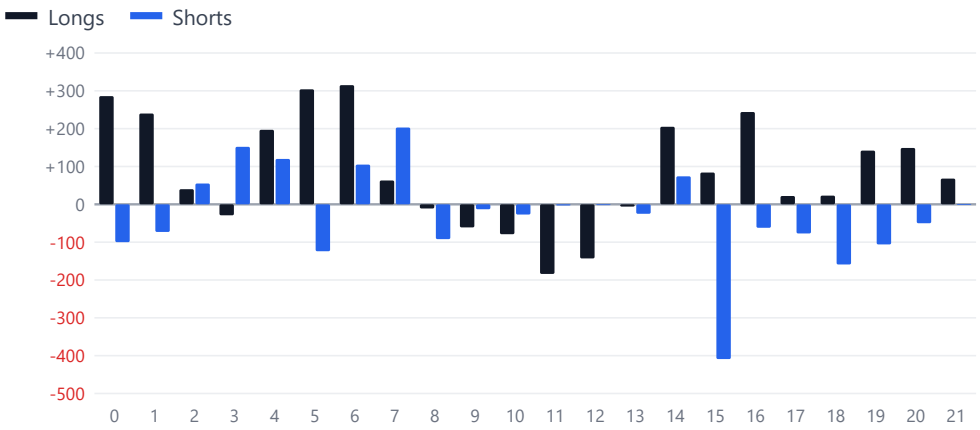
Same test as the 7-month study: every rule is known before the entry and simulated inside the engine. The 302 days were split in half; each rule's threshold was chosen on the first half (Jan 1 - Jul 31 2025) by the median improvement across 150 random parameter sets, then judged only on the second half (Aug 1 2025 - Feb 27 2026). The hour filter was also tested in New York time, since this period includes winter, when 10:00-13:00 New York is 15:00-18:00 UTC.

P&L by entry hour: average parameter set, first half vs second half



x: entry hour, UTC. y: net P&L of trades entered in that hour, pts, averaged over 150 random parameter sets. The losing hours moved: in the second half 8-12, 15 and 17-18 UTC lose (hour 18 for 98% of sets, hour 11 for 85%), while 14 UTC, a losing hour in 2026, was one of the best (+181).

P&L by entry hour: committed config.py, longs vs shorts



x: entry hour, UTC. y: net P&L, pts, Jan 2025 - Feb 2026. 8-13 UTC (London morning into the US open) is negative on both sides, mostly through the longs; shorts also lose heavily at 15 UTC. In 2026 the band negative on both sides was 14-16 UTC.

Filter results on the second half (out-of-sample)

Filter	First half: median set change	Committed config net	Committed config max DD	Median set net	Sets improved	Sets profitable	Trades kept
No filter	-	-293	1129	+285	-	80%	100%
Skip 10:00-13:00 New York	+176	-250	944	+257	38%	73%	87%
Skip 14:00-17:00 UTC (7-month study's version)	+95	-440	1076	+193	19%	66%	87%
Only trade with today's move	-70	-44	1016	+203	19%	71%	82%
Combined: skip 10-13 NY + with today's move (7-month winner)	+52	+29	684	+229	32%	76%	71%
Combined, UTC version	-47	-29	775	+251	39%	74%	71%
Each side only with its 10-day trend	-329	+457	727	+979	100%	100%	53%
Each side only with its 5-day trend	-1,149	+985	513	+1,103	100%	100%	55%
No new entries after day range > 80 pts	-93	-126	879	+543	83%	94%	84%
No new entries after day range > 30 pts	-665	+154	397	+544	75%	96%	36%
Stop after 4 losing trades per day	-261	+343	425	+572	81%	93%	56%
Daily loss stop -10 pts	-366	+231	400	+350	57%	81%	50%
No longs	-859	-1,032	1065	-613	1%	0%	42%

Second half, Aug 1 2025 - Feb 27 2026 (151 trading days), points. "First half: median set change" is what the filter did to the median set on Jan-Jul 2025, the data used to pick thresholds. No filter improved the median set on both halves: the trend gates that won on the second half were among the worst on the first, and the filters that looked good on the first half (the New York hour skip: +176) lost on the second.

Second-half P&L of 150 parameter sets: no filter vs the 7-month combined filter

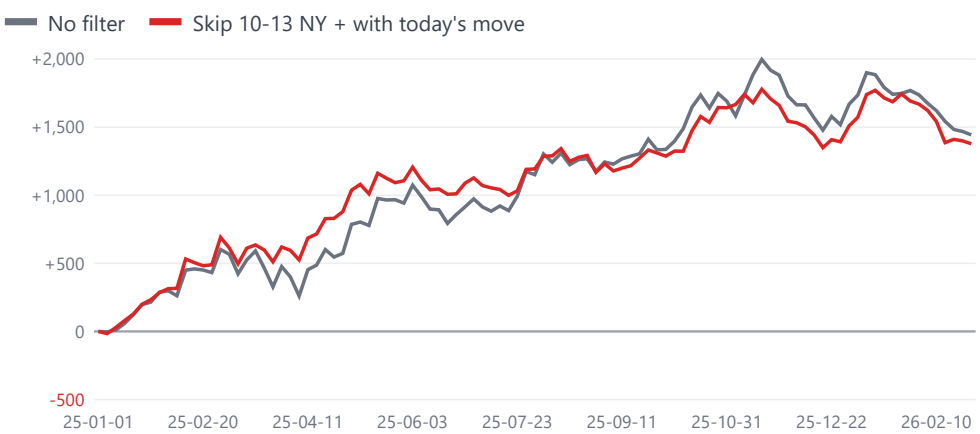
x: net P&L Aug 2025 - Feb 2026, pts (left edge of a 150-pt bin). y: number of parameter sets. No filter: 80% profitable, median +285. Combined filter: 76% profitable, median +229. In 2026 the same filter moved every set to the right.

Why it stopped working: "only trade with today's move" blocks longs whenever the price is below today's open, which in an uptrend includes many dips that were bought back. The hour skip follows the clock, but the losing hours here (8-12 UTC) are earlier than in 2026.

The trend gates ("each side only with its 5- or 10-day trend") are the mirror case: harmful in 2026 and in the first half of 2025, strongly positive in the steady uptrend from August 2025.

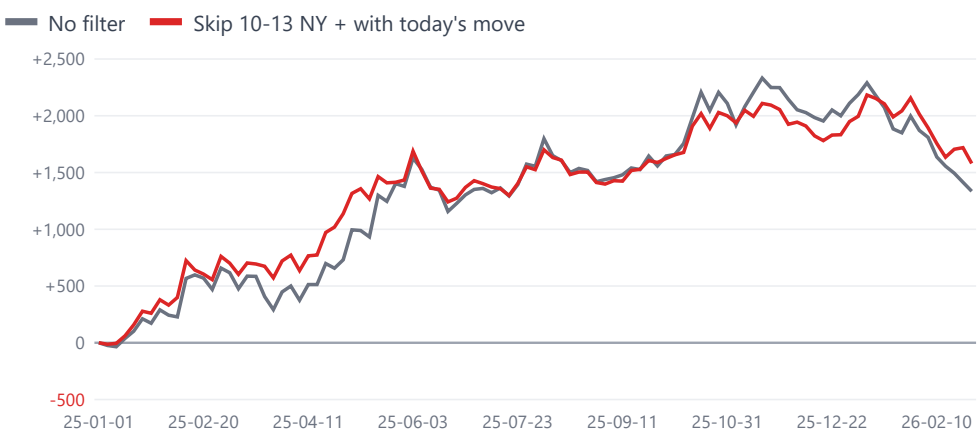
So every filter here is a bet on the regime. None is a structural fix like the 60-day window turned out to be.

Cumulative P&L, median of 150 parameter sets



x: trading day (YY-MM-DD, every 3rd day). y: median cumulative net P&L across the sets, pts. Split between first and second half on Aug 1 2025.

Cumulative P&L, committed config.py



x: trading day. y: cumulative net P&L, pts. Full period +1,262 without and +1,553 with the filter. The filter's main effect is a smaller drawdown (max DD 684 vs 1129 on the second half).

Monthly P&L with filters (points)

Filter	Jan 25	Feb 25	Mar 25	Apr 25	May 25	Jun 25	Jul 25	Aug 25	Sep 25	Oct 25	Nov 25	Dec 25	Jan 26	Feb 26
No filter	+292 / +306	+366 / +314	-209 / -132	+283 / +111	+655 / +386	-131 / -16	+299 / +226	-135 / +55	+184 / +163	+506 / +362	+14 / +13	-107 / -89	-38 / +126	-717 / -264

Filter	Jan 25	Feb 25	Mar 25	Apr 25	May 25	Jun 25	Jul 25	Aug 25	Sep 25	Oct 25	Nov 25	Dec 25	Jan 26	Feb 26
Combined: skip 10-13 NY + with today's move (7-month winner)	+364 / +301	+397 / +400	-39 / -64	+416 / +275	+304 / +225	-106 / +11	+190 / +57	-113 / +15	+201 / +152	+388 / +272	-102 / -76	+4 / -20	+202 / +161	-551 / -268
Each side only with its 10-day trend	+415 / +301	+368 / +211	+88 / +142	-14 / -28	+231 / +148	+12 / +25	+104 / +62	-99 / +66	+400 / +458	+580 / +420	-3 / +93	-102 / -85	-21 / +102	-297 / -84
No new entries after day range > 80 pts	+292 / +306	+366 / +314	-209 / -132	+327 / +93	+570 / +307	-120 / -4	+299 / +226	-135 / +55	+184 / +163	+523 / +561	-172 / -88	-48 / -66	-24 / +152	-453 / -277
Stop after 4 losing trades per day	+364 / +280	+295 / +290	-3 / +52	+166 / -58	+179 / +198	-28 / +8	+127 / +105	-110 / -43	+183 / +263	+413 / +406	+17 / -26	-9 / +108	+136 / +150	-288 / -264

Each cell: committed config.py / median of 150 random sets. For the median set the trend gate gives up most in Apr-Jul 2025 and gains in Mar and Sep 2025 and Feb 2026; the combined hour + direction filter is a little worse than no filter in most months.

5. Re-optimization and ML with the two filters switched on

All 3,001 parameter sets were re-simulated with the two 7-month filters (no entries 10:00-13:00 New York, only with today's move). The walk-forward re-chooses the parameters every 5 or 20 days on the best live-rule P&L of the previous 60 trading days. For the ML curves, every re-optimization also retrains the auto_pipeline.py-style ML filter on the chosen set's filtered trades over the same 60 days (logistic regression, random forest, gradient boosting and soft voting, common and long/short-specific; 70/30 split, gate at least 30 trades and PF 1.3).

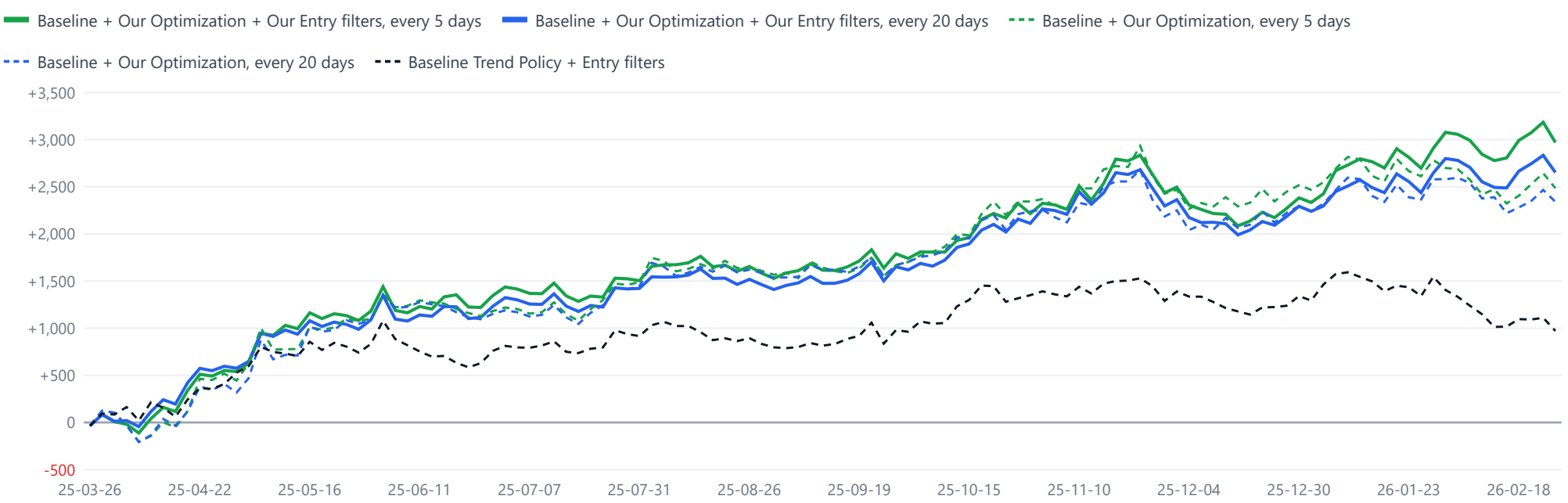
+2,997 pts
Every 5 days on 60 days, with filters (+2,610 without); better in 5 of 5 phases

+2,658 pts
Every 20 days, with filters (+2,448 without); better in only 11 of 20 phases

-1,279 pts
Profit cost of our ML, every 5 days, for a max drawdown of 310 instead of 774

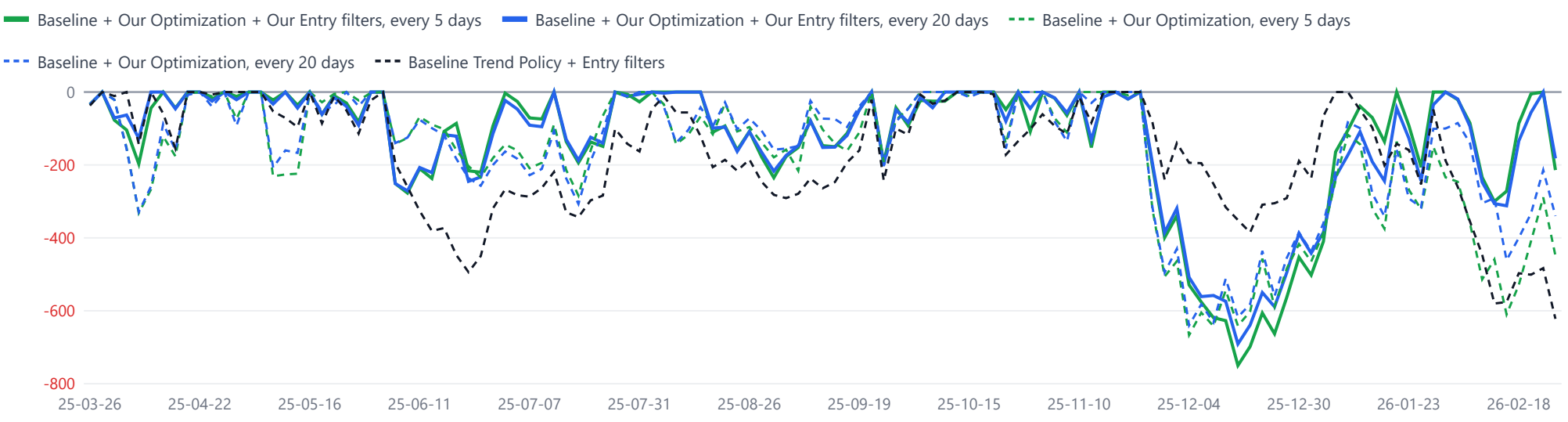
-884 pts
Profit cost of our ML, every 20 days, for a max drawdown of 322 instead of 719

Cumulative P&L: re-optimization with and without the filters



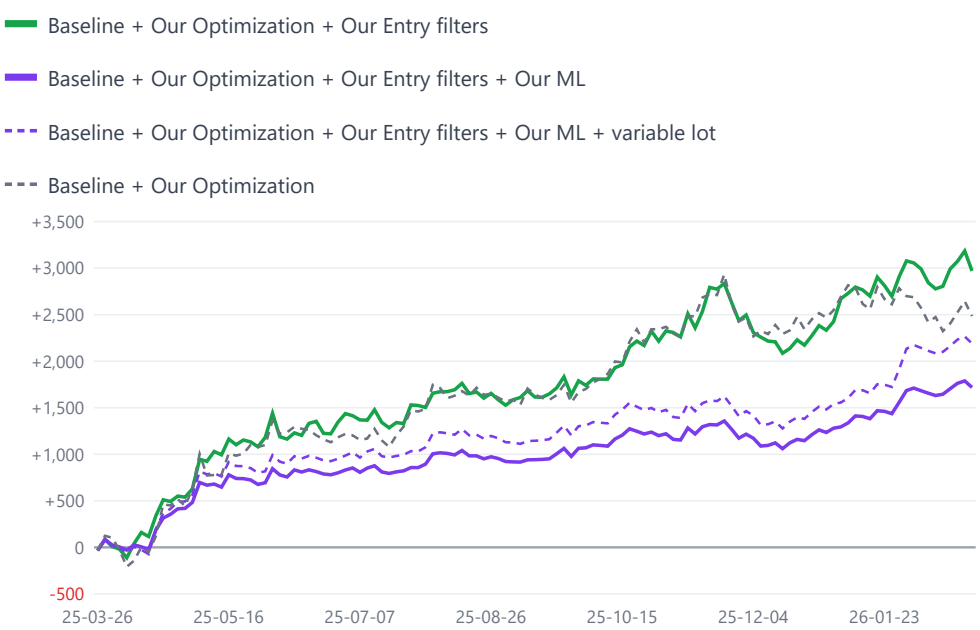
x: trading day (YY-MM-DD). y: cumulative net P&L, points, averaged over every re-optimization phase. Out-of-sample test: 2025-03-26 to 2026-02-27 (242 trading days); the first 60 days are training-only. The entry filters help mostly in Jan-Feb 2026. Applied to fixed sets they lower the median from +1,176 to +936; the Baseline Trend Policy goes from +923 to +944 with a smaller drawdown (684 vs 1129). Baseline Trend Policy = committed config.py with fixed parameters; every other line is a SAIL Lab proposal built on it. Optimization = walk-forward re-optimization on a 60-day window. Entry filters = no entries 10:00-13:00 New York (US/Europe overlap) and only in the direction of today's move.

Drawdown: re-optimization with and without the filters



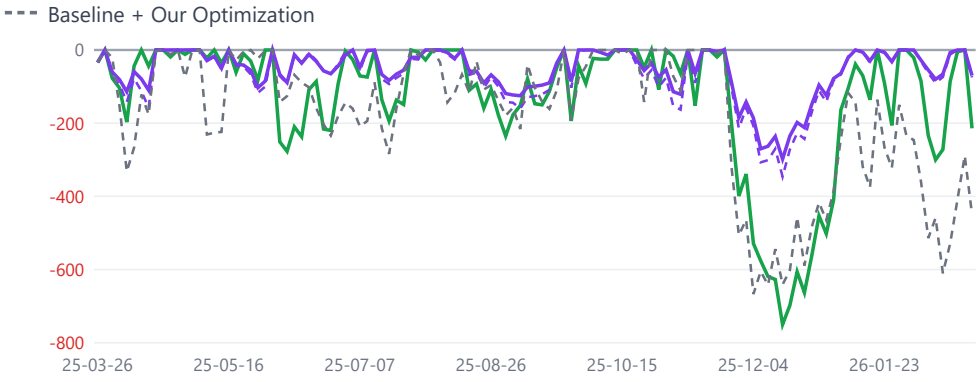
x: trading day (YY-MM-DD). y: drawdown, points (cumulative P&L minus its running peak; 0 = new high). Deepest: every 5 days -774 with entry filters vs -667 without; every 20 days -715 vs -645; Baseline Trend Policy + Entry filters -684. Drawdown of the phase-averaged curve; deepest values are exact daily figures (the chart shows every 2nd day). The summary table's max DD is the average of each phase's own drawdown.

Re-optimization + best ML + filters, every 5 days



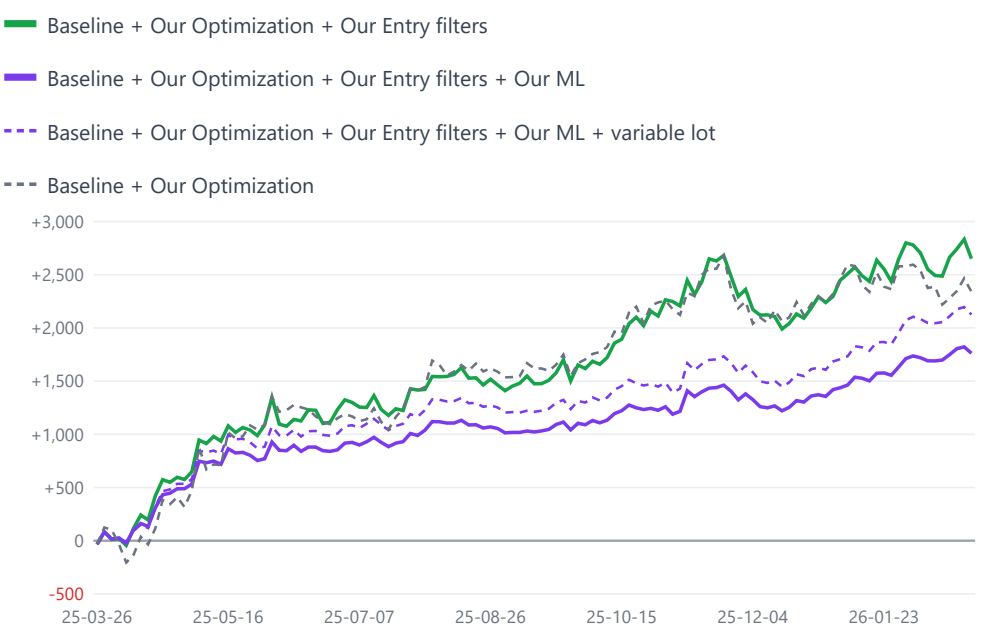
x: trading day. y: cumulative net P&L, points (variable lot: P&L x proba/threshold, clipped to 1-2x). A model was deployed at 77% of 246 retrains. ML cut the drawdown to 310 but gave up +1,279 pts.

Drawdown, every 5 days



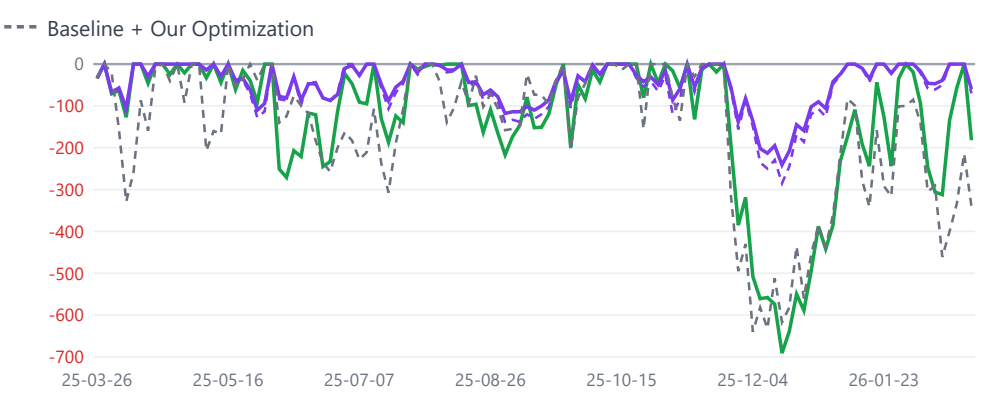
x: trading day. y: drawdown, points (0 = new high). Deepest: + Entry filters -774, + ML -298, + variable lot -346, Optimization alone -667. Drawdown of the phase-averaged curve; deepest values are exact daily figures (the chart shows every 2nd day). The summary table's max DD is the average of each phase's own drawdown.

Re-optimization + best ML + filters, every 20 days



x: trading day. y: cumulative net P&L, points. Deployed at 72% of 261 retrains. The trades the ML skipped were net profitable, but its drawdown is lower than random trade dropping at the same P&L (see the control test below).

Drawdown, every 20 days



x: trading day. y: drawdown, points (0 = new high). Deepest: + Entry filters -715, + ML -241, + variable lot -286, Optimization alone -645. Drawdown of the phase-averaged curve; deepest values are exact daily figures (the chart shows every 2nd day). The summary table's max DD is the average of each phase's own drawdown.

Every schedule, without vs with the filters

Score	Window	Daily	Every 5 d	Every 10 d	Every 20 d
Current pipeline score	20 days	+1,614, then +926	+1,672, then +956	+1,490, then +796	+1,662, then +949
Current pipeline score	40 days	+500, then +478	+1,378, then +611	+1,531, then +746	+1,678, then +888
Current pipeline score	60 days	+1,434, then +1,040	+1,090, then +916	+1,133, then +1,050	+1,153, then +1,021
Net P&L, live rules	20 days	+1,374, then +2,430	+1,904, then +2,651	+1,718, then +2,317	+1,734, then +2,155
Net P&L, live rules	40 days	+1,902, then +2,552	+1,991, then +2,149	+2,137, then +2,370	+2,044, then +2,383
Net P&L, live rules	60 days	+2,486, then +2,933	+2,610, then +2,997	+2,482, then +2,895	+2,448, then +2,658
Plateau (30 nearest sets)	20 days	+2,484, then +2,712	+2,379, then +2,688	+2,121, then +2,503	+2,077, then +2,266
Plateau (30 nearest sets)	40 days	+2,413, then +2,684	+2,362, then +2,585	+2,713, then +2,546	+2,722, then +2,518
Plateau (30 nearest sets)	60 days	+2,669, then +2,929	+2,904, then +2,802	+2,833, then +2,887	+2,844, then +2,782

Out-of-sample net pts Mar 26 2025 - Feb 27 2026, averaged over phases. Each cell: without filters, then with filters. The filters improved 20 of 36 schedules: all 12 live-P&L schedules and 8 of 12 plateau ones, but none of the 12 pipeline-score ones (in 2026 they improved all 36).

Summary

Strategy	Net	Max DD	Phases	Better in
Baseline + Our Optimization (Every 5 days)	+2,610	702	+2,414 to +2,794	-
Baseline + Our Optimization + Our Entry filters (Every 5 days)	+2,997	774	+2,814 to +3,145	5 of 5
Baseline + Our Optimization + Our Entry filters + Our ML (Every 5 days)	+1,718	310	+1,526 to +2,188	0 of 5
Baseline + Our Optimization + Our Entry filters + Our ML + variable lot 1-2x (Every 5 days)	+2,184	359	+1,832 to +2,758	0 of 5
Baseline + Our Optimization (Every 20 days)	+2,448	709	+1,614 to +3,348	-
Baseline + Our Optimization + Our Entry filters (Every 20 days)	+2,658	719	+1,984 to +3,482	11 of 20
Baseline + Our Optimization + Our Entry filters + Our ML (Every 20 days)	+1,774	322	+1,054 to +2,475	1 of 20
Baseline + Our Optimization + Our Entry filters + Our ML + variable lot 1-2x (Every 20 days)	+2,137	392	+1,176 to +3,007	6 of 20
Baseline Trend Policy, without / with Entry filters	+923 / +944	1129 / 684	-	-
Median of 3,001 fixed sets, without / with Entry filters	+1,176 / +936	-	-	-

"Better in": phases where the row beat the line above it (+ Entry filters vs Optimization alone; + ML rows vs + Entry filters). Baseline Trend Policy = committed config.py with fixed parameters; every other line is a SAIL Lab proposal built on it. Optimization = walk-forward re-optimization on a 60-day window. Entry filters = no entries 10:00-13:00 New York (US/Europe overlap) and only in the direction of today's move.

Does ML lower drawdown by skill, or only by trading less?

Control test, no retraining: take the trades of "Baseline + Optimization + Entry filters" in every phase (rebuilt from the sweep, within 0.05 pts of the study's phase totals), keep each trade at random with probability = (P&L with ML) / (P&L without ML) so the expected P&L equals ML's, and repeat 1,000 times. If ML's drawdown is no smaller than random dropping, its lower risk is only the effect of trading less. Drawdown here is the per-phase max DD averaged over phases, as in the summary tables. W = 60 days; the 7-month and last-14-month studies are included for comparison.

12-38%
Lower drawdown than random dropping at the same P&L, in all 5 testable schedules

0-13.5%
Share of random draws with a drawdown as small as ML's (4 of 5 below 5%)

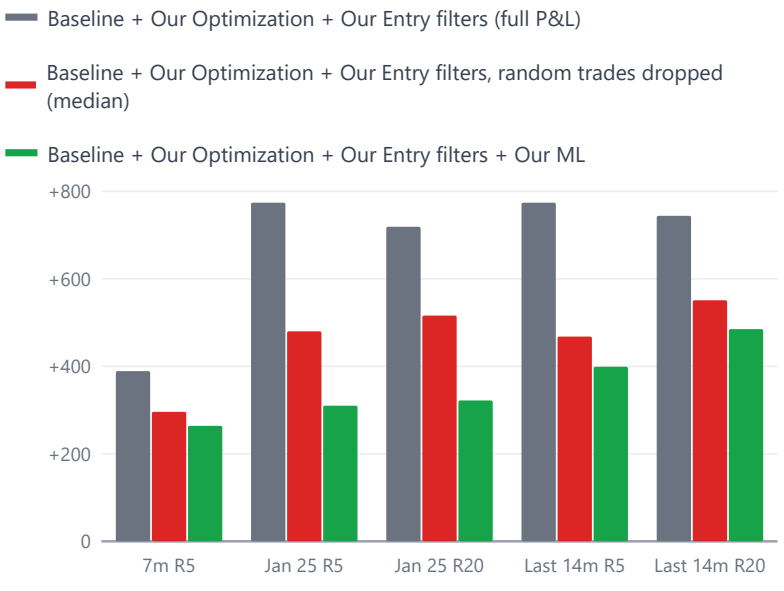
33-51%
P&L given up by adding ML to Optimization + Entry filters, in 5 of 6 schedules: ML controls risk, it does not add return

Same P&L, different drawdown: ML vs random trade dropping

Schedule	Kept	Opt. + Entry filters	Opt. + Entry filters + ML	Random dropping (median)	Random DD 5-95%	Random DD <= ML	Random ratio >= ML
7 months, every 5 d	57%	+1,165 / 389	+667 / 264	+674 / 296	250-346	13.5%	33.3%
7 months, every 20 d	-	+903 / 345	+932 / 254	-	-	-	-
14 months Jan 25, every 5 d	57%	+2,997 / 774	+1,718 / 310	+1,707 / 480	409-548	0%	0.2%
14 months Jan 25, every 20 d	67%	+2,658 / 719	+1,774 / 322	+1,777 / 516	480-551	0%	0%
Last 14 months, every 5 d	49%	+2,492 / 774	+1,216 / 399	+1,227 / 468	400-549	4.9%	23.9%
Last 14 months, every 20 d	56%	+1,788 / 744	+996 / 485	+998 / 551	503-608	1.2%	17.8%

Cells: net pts / max DD. "Random DD <= ML" and "Random ratio >= ML": share of the 1,000 draws that matched or beat ML; low = ML's edge is not luck. 7 months every 20 days cannot be tested: adding ML made more (+932 vs +903) with a smaller drawdown, so no random subset can match its P&L; ML is simply better there. All columns are Baseline Trend Policy + Optimization (60-day window) + Entry filters, with or without ML.

Max drawdown at the same P&L



x: study and re-optimization interval (R5 = every 5 days). y: max drawdown, points. Random dropping and ML have the same expected P&L; the version without ML has the full, higher P&L.

Reading

Dropping trades at random also lowers drawdown, but less than P&L, so P&L / DD gets worse (7 months every 5 days: Optimization + Entry filters 2.99, random 2.27). ML lowers drawdown well beyond random dropping in every testable schedule, so it does pick out losing trades. The edge in P&L / DD is strong on the 14 months from Jan 2025 (random almost never matches it) but weaker on the 7-month every-5-day and last-14-month runs (18-33% of random draws match it). ML is a risk control, not a return booster.

---

6. Longs vs shorts, with gold as the benchmark

Same walk-forward schedules as section 5, split by trade direction. The gold line is the change in XAUUSD's daily closing bid since the Mar 25 2025 close (3,021.07), in the same points: what holding one unit of gold long would have made. The strategy can hold up to 4 positions at once, so its exposure is not fixed at one unit.

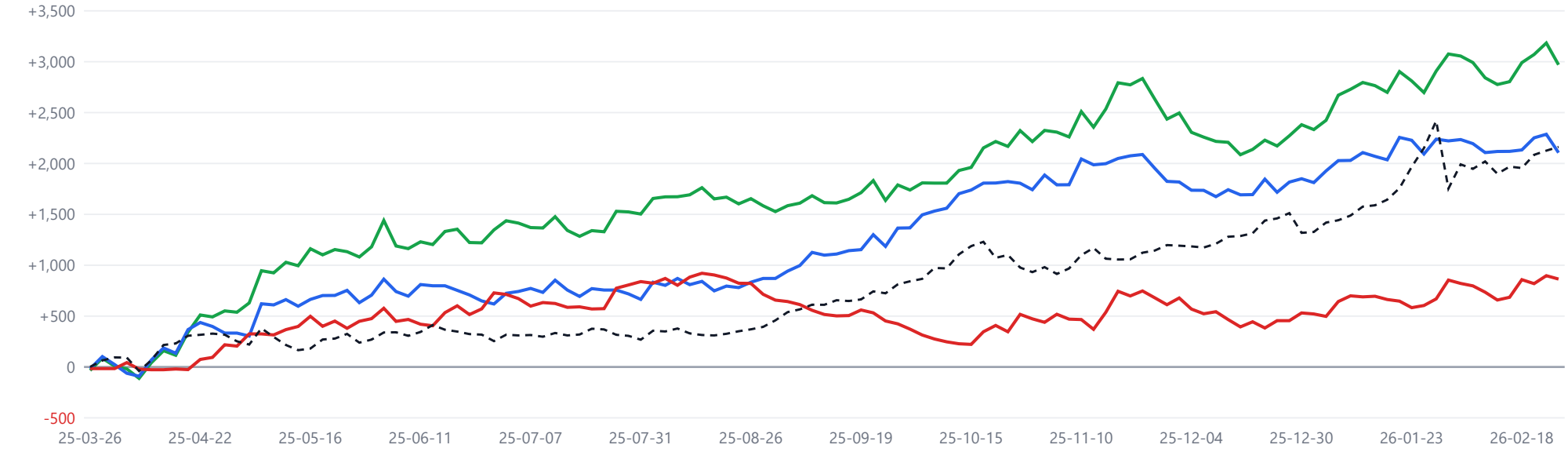
+2,181 pts
Longs, re-optimized every 5 days with filters: 1433 trades. Without filters: +2,907

+816 pts
Shorts, same schedule: 1024 trades. Without filters: -298

+2,163 pts
Gold over the same period (+75%, 3,021 to 5,279). Buy-and-hold made this; short-and-hold lost it

+2,997 pts
Strategy total with filters, 1.4x buy-and-hold (in 2026: 4.6x short-and-hold)

Cumulative P&L by direction vs gold: re-optimize every 5 days on 60 days, with filters



x: trading day (YY-MM-DD). y: cumulative net P&L, points (gold: price change in points). Averaged over the 5 re-optimization phases. Out-of-sample test: 2025-03-26 to 2026-02-27 (242 trading days); the first 60 days are training-only. The longs track gold. The shorts made money in the quieter months (Apr-Jul 2025) and lost in the Aug-Sep 2025 rally; in Jan 2026 both sides made money.

Every 20 days on 60 days, with filters

x: trading day. y: cumulative net P&L, points. Longs +2,085 (1410 trades), shorts +573 (1015 trades). Without the filters: longs +2,765, shorts -317.

Every 5 days on 60 days, without filters

x: trading day. y: cumulative net P&L, points. Longs +2,907 (2064 trades), shorts -298 (1560 trades). The filters cost the longs -726 and saved the shorts +1,114: "with today's move" blocks counter-trend shorts on up days.

Monthly P&L by direction vs gold's move (points)

Month	Gold move	Longs, filters	Shorts, filters	Longs, no filter	Shorts, no filter
Mar 25 (26-31)	+100	+114	-16	+171	-25
Apr 25	+152	+219	+222	+133	+165
May 25	+17	+372	+270	+420	+235
Jun 25	+23	+6	+238	-23	+108
Jul 25	-24	-47	+126	+124	+180
Aug 25	+159	+219	-182	+234	-156
Sep 25	+415	+509	-270	+499	-394
Oct 25	+139	+494	+52	+505	+192
Nov 25	+218	-62	+172	+372	-314
Dec 25	+100	-12	-90	-2	+42
Jan 26	+573	+427	+208	+383	-94
Feb 26	+388	-57	+86	+92	-239

Re-optimized every 5 days on 60 days, averaged over phases. Gold move = last close of the month minus last close of the previous month. Green = gold rose more than 300 pts, red = gold fell.

The two periods together answer the open question from the 7-month study: the strategy has no structural long or short bias. It makes money on the side of the trend, and the 60-day re-optimization follows the trend side automatically.

Running short-only, which the 7-month data suggested, would have lost heavily here: the no-filter shorts were negative, and "no longs" cut the second-half P&L of 99% of the parameter sets (section 4).

What to do

1. **Re-optimize on a 60-day window, scored on live-rule P&L or plateau.** It is the only result that held in both a falling and a rising market (beats 93-98% of fixed sets here; with live-P&L scoring 95-100% in 2026). The interval can be anywhere from 5 to 20 days. Stop using the pipeline score (calculate_score on tick_engine): it was a coin flip in both periods.
2. **Do not hard-code the two filters as permanent rules.** They were the best result of the 7-month study, but here they improved only a third of the parameter sets and lowered the median fixed set. Inside the 60-day re-optimization they were neutral to slightly positive (+387 every 5 days, +210 every 20 days). If you keep them, treat them as a bear-market setting.
3. **Trade both sides.** Do not switch to short-only: longs made the money in 2025 just as shorts did in 2026.
4. **Turn our ML on when risk matters more than return.** On top of re-optimization it cut the max drawdown by 55-60% (774 to 310 every 5 days), well beyond random trade dropping, for -1,279 pts of profit. For the highest profit, run it without ML.
5. **Next test:** run the walk-forward over the full 21 months (Jan 2025 - Oct 2026) in one piece, to measure the 60-day schedule through the change from bull to bear market, and compare a regime-aware filter (for example the 60-day daily trend from section 3) against no filter.

Reproduce (RESEARCH_DATA=14m, RESEARCH_CONFIG=committed config.py, FILTER_TZ=ny): USE_OLD_PARQUET=0 build_cache_12m.py 2025-01-01 2026-02-28, param_sweep.py 3000 20, SWEEP_FILTERS=1 param_sweep.py 3000 20, walkforward_12m.py (WF_START=60/40/120), regime_attribution.py, day_filters.py, filters_study.py 150 16, long_short_gold.py, ml_window_study.py, opt_filters_ml.py 16, canvas_data_14m.py, canvas_ts_14m.py. Runner: output\research\14m\run_all.ps1. Outputs: output\research\14m\.