

$\text{eval}^2 \propto d^2$:

evaluating evals before they
dictate deployment decisions

Russell Ro
October 1, 2026

What's an eval? What's a benchmark?

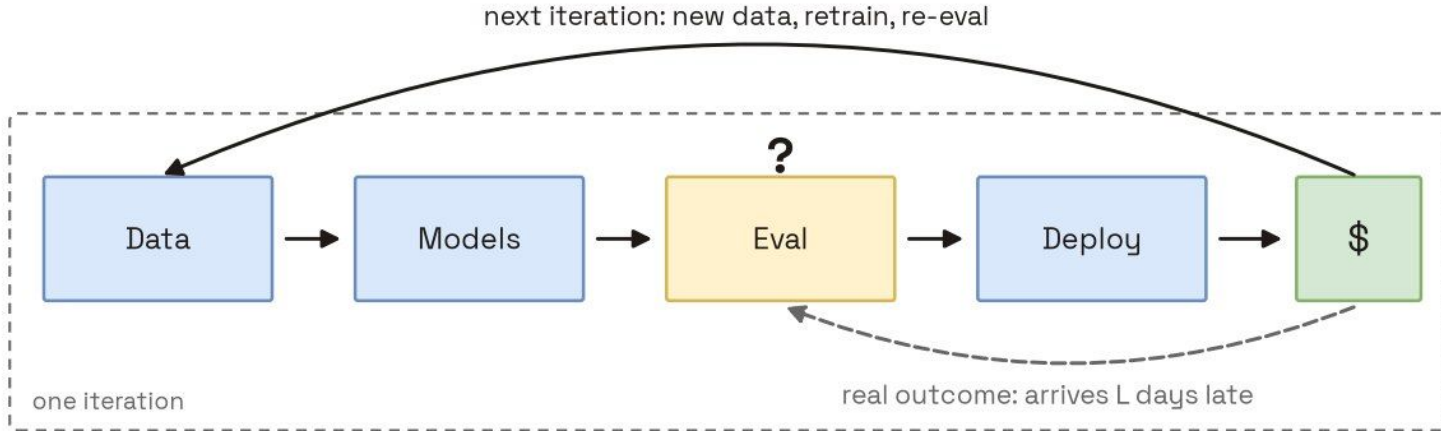
Eval: measuring how a system does on a task, using one or more metrics. Can be per example, per task, or whole system in deployment. [\[1\]](#)[\[2\]](#)

Benchmark: a fixed dataset + metric, shared so people can compare methods. [\[3\]](#)

So every benchmark is an eval, but most evals aren't benchmarks. [\[1\]](#)[\[3\]](#)

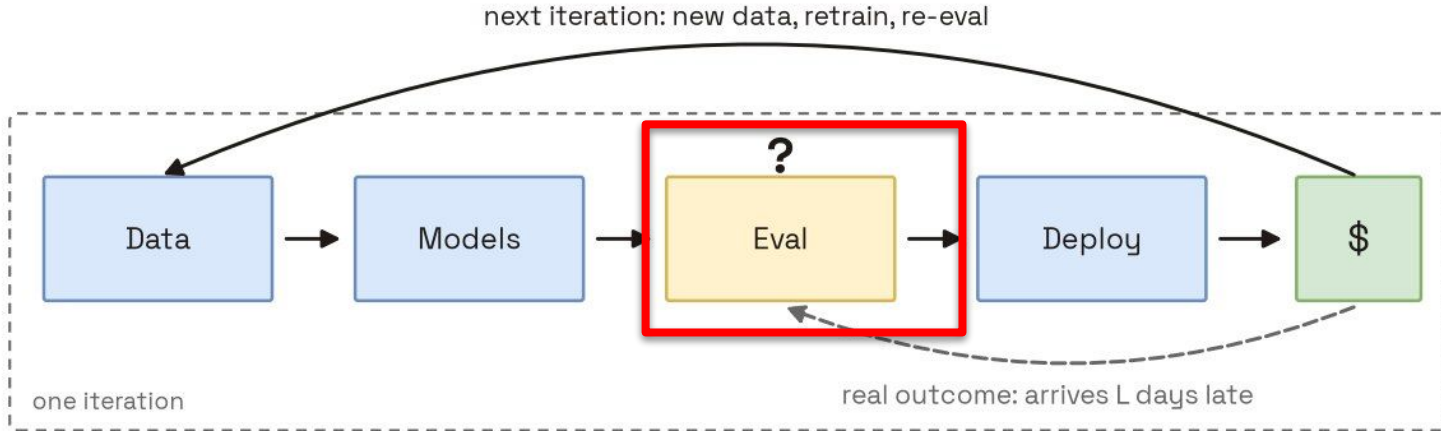
Either way it's a proxy for what you care about (\$\$\$), and it only counts while it keeps predicting \$\$\$\$. [\[4\]](#)[\[5\]](#)

ML lifecycle



- Prep the data, train the models, evaluate the models, use eval results to drive deployment decisions, wait, ???, \$\$\$, optimize, repeat.

ML lifecycle



- Prep the data, train the models, evaluate the models, use eval results to drive deployment decisions, wait, ???, \$\$\$, optimize, repeat.

Data Distribution Shifts / Drift

$$P(X, Y) = P(Y | X) P(X) = P(X | Y) P(Y)$$

Covariate (Data) Shift: $P(x)$ moves, $P(y|x)$ holds

Label Shift: $P(y)$ moves, $P(x|y)$ holds

Concept Shift: $P(y|x)$ moves, $P(x)$ holds

The above are metrics (evals) that are generally used for ML monitoring.

This includes **evaluating our model selection evals**, especially if evals sit as a conduit between the models and the money that they make.

Discussion question: What type of shift would it be if our benchmarks start becoming a poor proxy for how predictive a model is at making money?

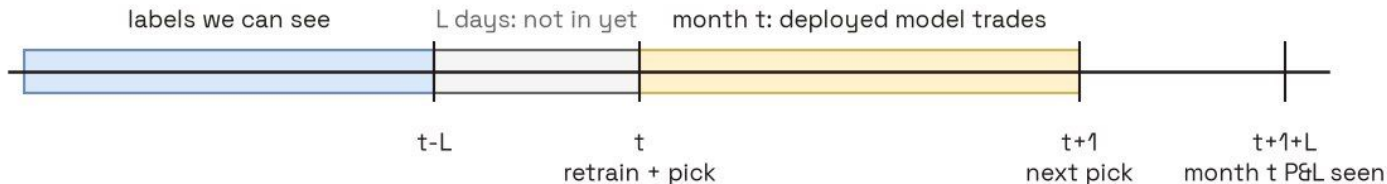
Worked example: SPY trading, *but worse*

Setup: SPY daily, 2010–2025. Target is **next-day up/down**.
Features are 1/5/20-day returns, MA10/MA50 gaps, 20-day vol, and day/month.

Model zoo (8): LogReg, LogReg-mom, RF, ExtraTrees-vol, GBM, GBM-mom, kNN, MLP (each gets its own feature subset)

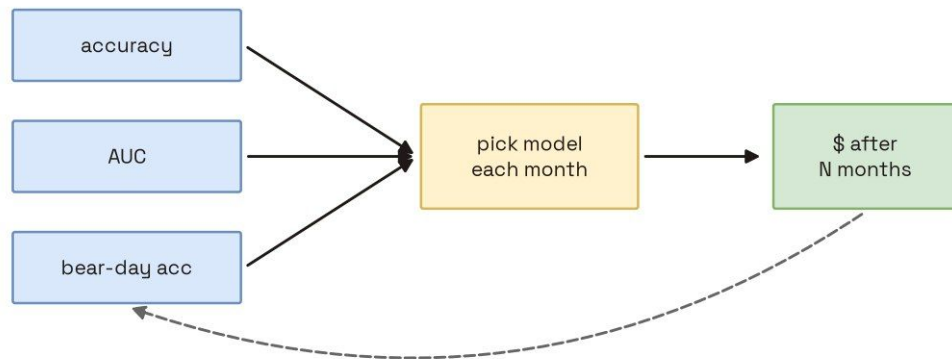
Monthly loop:

1. At month t , retrain each model and score it on data through $t-L$ ($L = 21$ trading days)
2. The eval metric picks one model
3. Deploy it for the month
4. The P&L only shows up L days later



Evaluating the Eval

- Because of the L-day lag (for both observing data and making actions), we can't just deploy the model that made money in short time frames. We have to pick on a **proxy metric** (or combination of metrics).
- Treat each metric as a selection rule then run through the monthly loop.
- Grade each rule by the \$ it earned from 2018 to 2025.



Metrics:

Accuracy, AUC, Negative log loss, Bear accuracy, Lagged P&L, Combined

Metrics

Selection rules:

- **Accuracy:** Share of up/down calls that were correct: $(1/n) \sum 1[\hat{y} = y]$.
- **AUC:** How well probabilities rank up days above down days: $P(p_{\text{up}} > p_{\text{down}})$; 0.5 = random, 1 = perfect.
- **Negative log loss:** Rewards calibrated probabilities and penalizes confident errors: $(1/n) \sum [y \log p + (1-y) \log(1-p)]$; higher is better.
- **Bear accuracy:** Accuracy when SPY is below its 50-day MA; if the window has fewer than 3 such days, keep last month's pick.
- **Lagged P&L:** Sum of daily long/short returns: $\sum s \cdot r$, where $s = +1$ for predicted up and -1 for predicted down.
- **Combined:** Average of the model's percentile ranks across the five rules above for that month.

References (not selection rules):

- **Never switch:** Keep the model selected by accuracy in month one for the entire run.
- **Always long:** Buy SPY once and hold it throughout the run.
- **Oracle:** Select the best model each month with hindsight (an upper bound).

Three heuristics to consider

- **H1. Validity:**

Does the selector score track \$?

- **H2. Temporal holdout:**

Does the best selector survive over time?

- **H3. Cross-eval comparison:**

What do the evals say about each other?

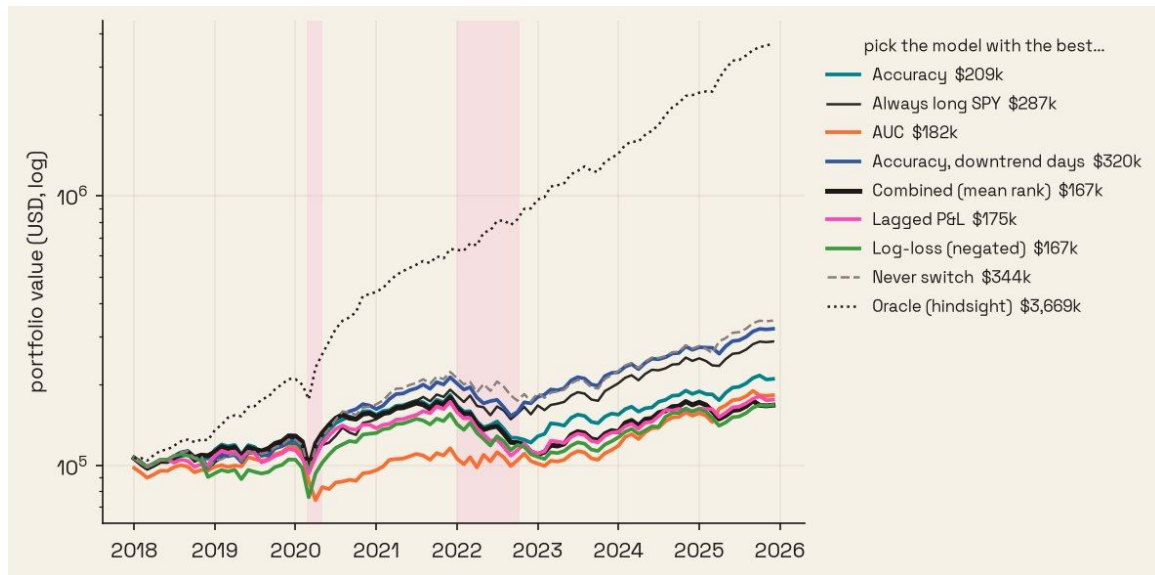
H1. Can the score track \$?



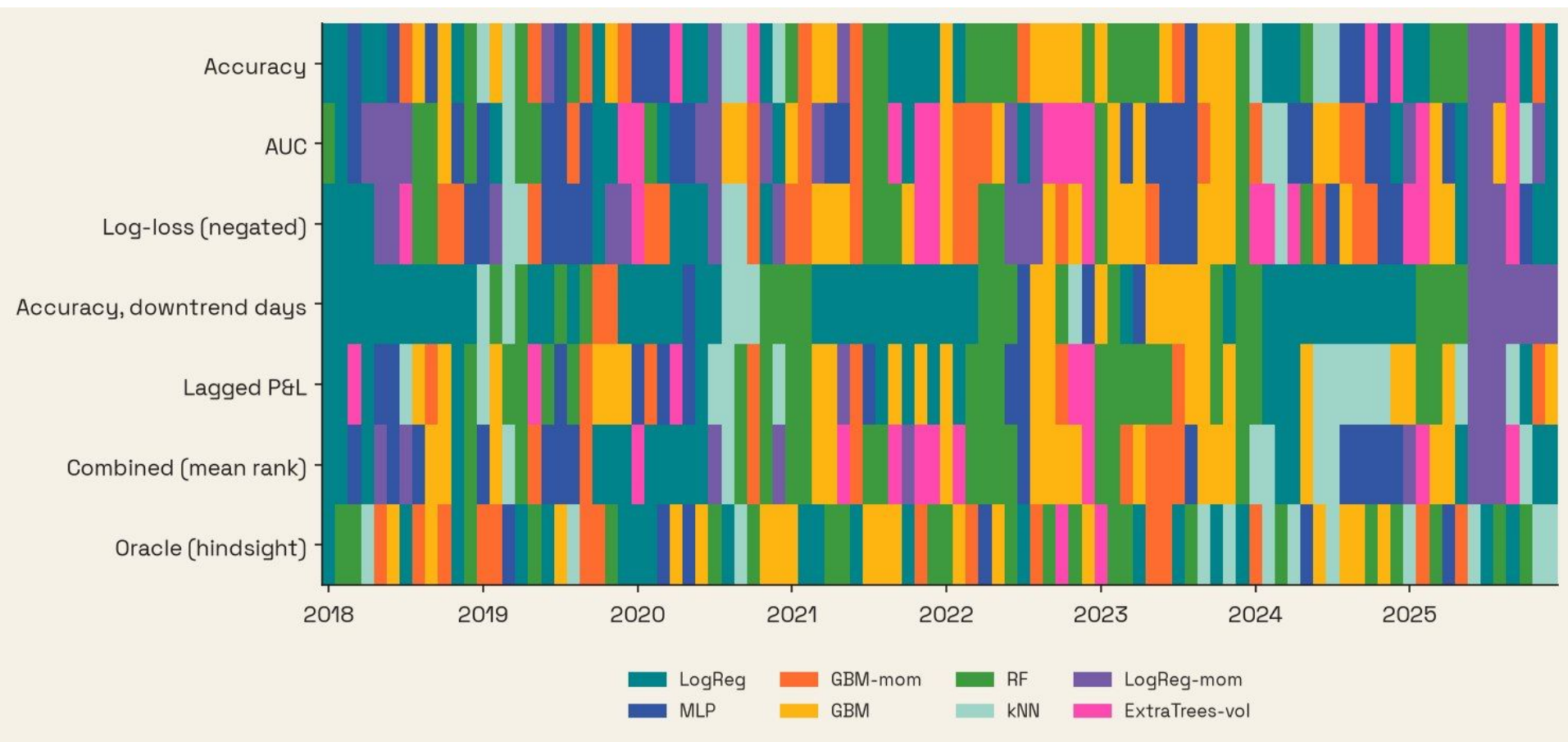
- Each month: correlate each metric's score with that month's Sharpe
- Accuracy: corr 0.82, stays above 0.62 in every rolling window
- AUC: corr 0.35, LogReg falls to -0.55 in 2023
- A metric can look fine overall and still flip sign when the regime changes

H1. Can the score track \$?

- \$100k deployed by each rule, 2018–2025, net of fees + short borrow
- **Bear-day acc \$320k** · Accuracy \$209k · AUC \$182k · Log-loss \$167k
- **Never switching \$344k** · **SPY \$287k**
- Fees cost switchers \$10-15k (60+ swaps). At 10bp only bear-day acc (\$220k) and never-switch (\$288k) stay near SPY

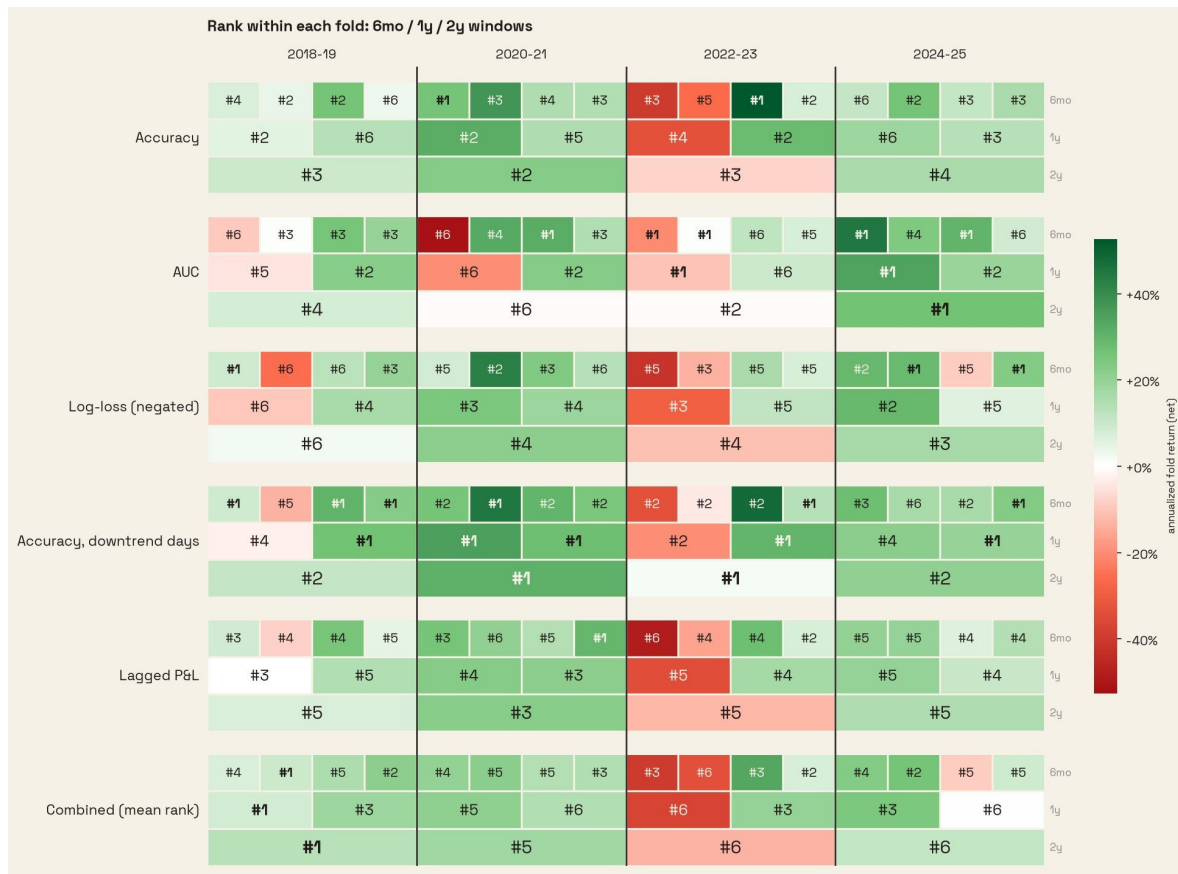


Best correlation $\neq$ most money

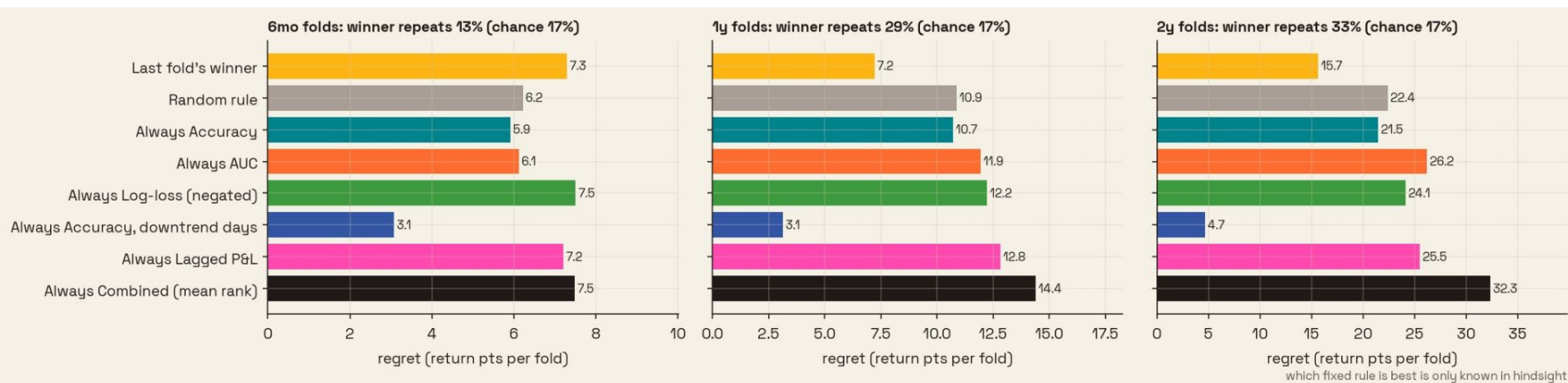


H2. Does the best selector survive time?

- Blocked folds of 6mo / 1y / 2y, not shuffled.
- Last fold's winner repeats 13% of the time at 6mo (chance 17%), 29% at 1y.
- Bear-day accuracy steadies as the window grows: avg rank 2.2 $\rightarrow$ 1.9 $\rightarrow$ 1.5.



H2. Does the best selector survive time?



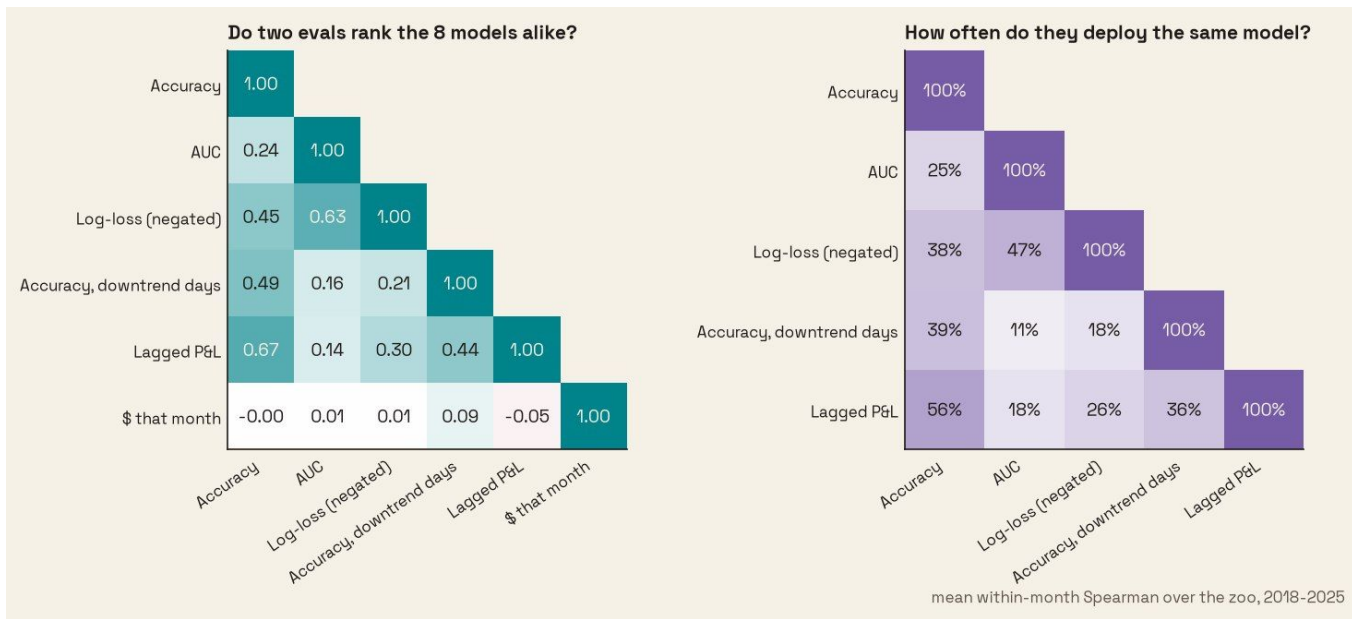
- Regret = return points lost per fold vs that fold's best rule in hindsight (lower = better).
- Last fold's winner: 7.3 / 7.2 / 15.7 pts (6mo / 1y / 2y), worse than a random rule at 6mo.
- Always bear-day accuracy: 3.1 / 3.1 / 4.7.

H3. What do the evals say about each other?

Short recap:

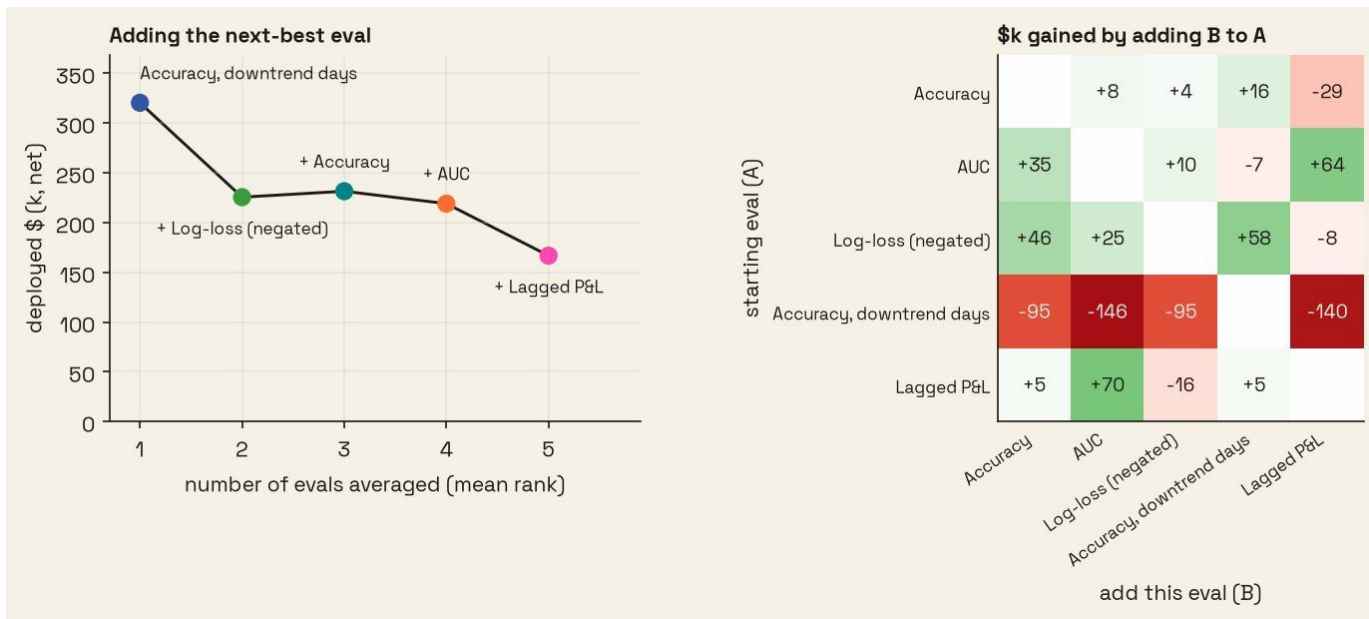
- **H1:** does each eval signal \$?
- **H2:** does it hold over time?
- **H3:** what do they add together?
- An eval suite is a portfolio:
 - Correlated evals may add the same information as one another
 - Similarly, decorrelated evals may add information to our understanding of the problem

H3. What do the evals say about each other?



- Within-month rank corr over 8 models, averaged over 96 months (2018-25).
- Two clusters: right-or-wrong (accuracy + lagged P&L, 0.67) and calibration (AUC + log-loss, 0.63). Across clusters: 0.24.
- Only bear-day accuracy tracks that month's \$ (0.09). The rest sit within ± 0.05 .

H3. What do the evals say about each other?

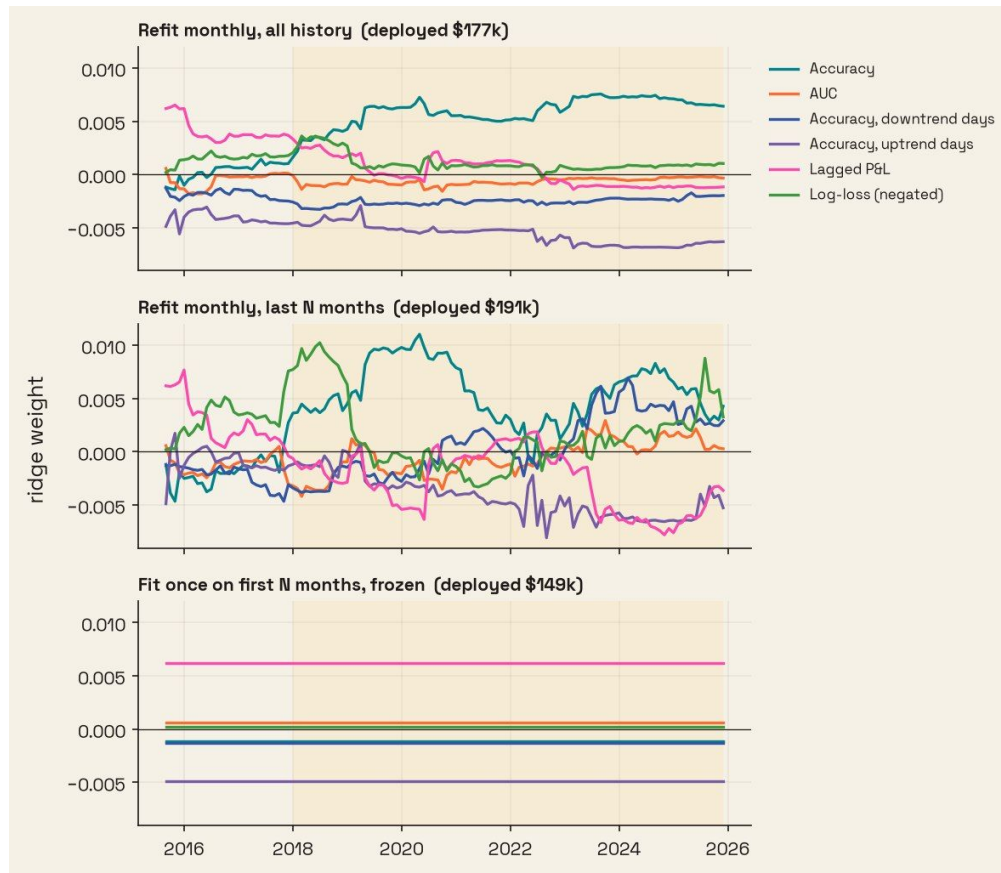


- Start with the best eval, keep adding whichever helps most. Combined = mean rank.
- Bear-day acc alone \$320k. +log-loss \$226k, +acc \$232k, +AUC \$219k, all five \$167k (net).
- Every add costs money: averaging dilutes the one eval that carries \$ info.

H3. What do the evals say about each other?

Can we learn the selector?

- Ridge on the six eval signals: learn which ones predict next month's \$.
- Frozen on 2018-21, deployed 2022-25: \$120k, 4th of 8. Below AUC (\$157k), bear-day accuracy and always long (\$151k).
- Rank IC about 0. Weights that worked before stop working after: drift breaks a static selector.
- **It frankly sucks, but it's a neat idea.**



Extra ideas: Where might learned selectors work?

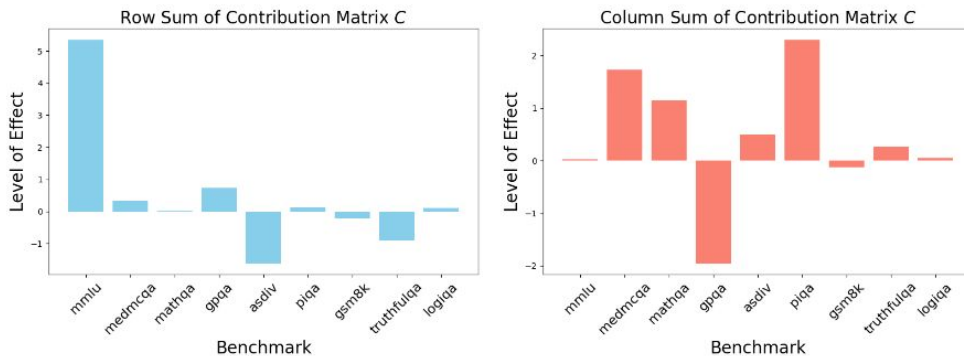
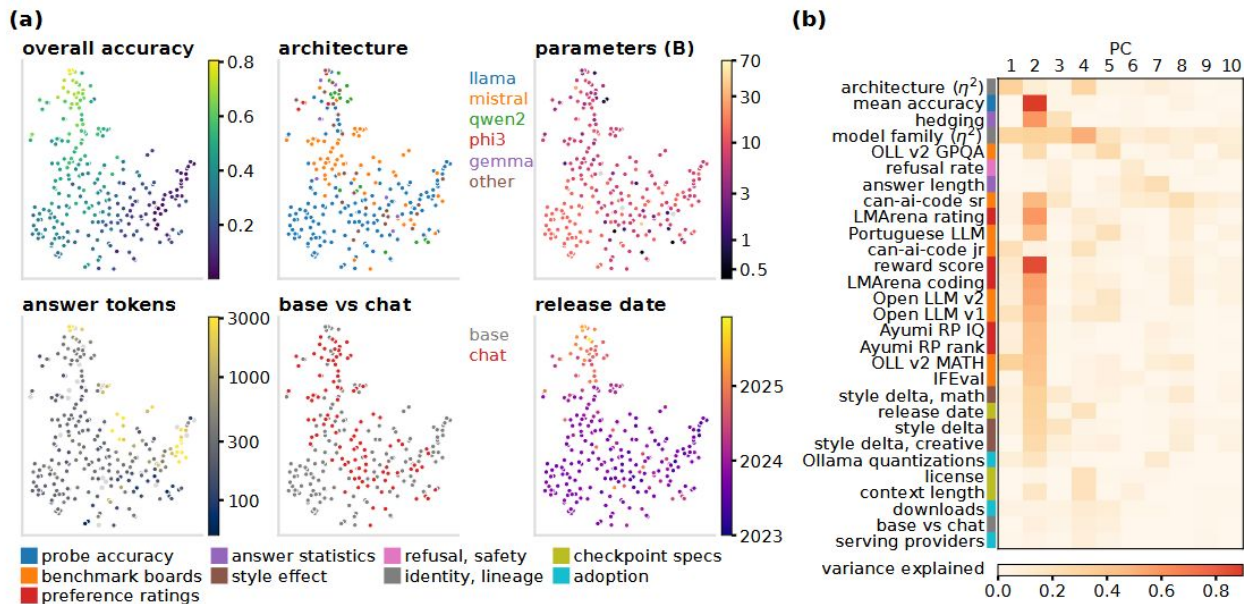


Figure 7: **Left:** Effect of removing benchmarks on testing all other benchmarks. Higher value suggests that the addition of that benchmark into the training of model embedding enhances predicting model correctness on the rest of the benchmarks. **Right:** Effect of removing each of the rest of the all other benchmarks on the benchmark being tested. Higher value suggests that predicting models' correctness in answering questions from the tested benchmark becomes easier with each of the rest of the benchmarks added overall.

I got the learned selector idea from [EmbedLLM](#): given a model's performance on a benchmark, can you predict it on other benchmarks (short answer: for LLMs, yes).

Fig: Zhuang, *et al.*

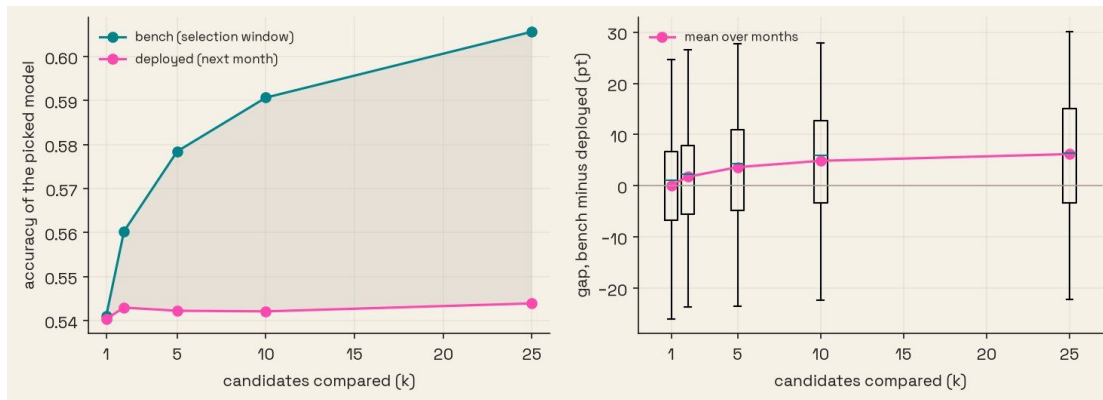
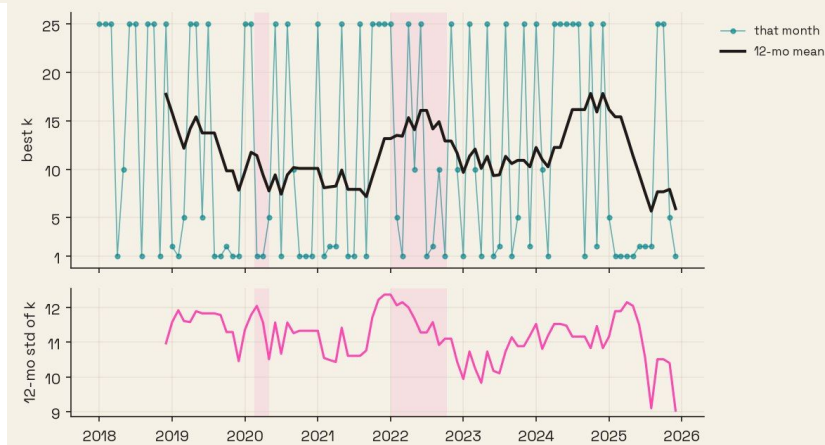
Extra ideas: Where might learned selectors work?



I got the learned selector idea from [EmbedLLM](#): given a model's performance on a benchmark, can you predict it on other benchmarks (short answer: for LLMs, yes).

Fig: Ro, *et al.*

Extra ideas: more models $\neq$ better



- Best of k GBM configs: eval climbs 54% $\rightarrow$ 61% as k grows. Deployed stays ~54% for every k.
- Eval overpromises by 6.2pt at k=25, 0.1pt at k=1.
- Why: within a month, eval vs deployed acc correlates 0.02. The k=25 winner is just the luckiest config on a 63-day window.

Insights

- Predictive: plain accuracy (0.82 corr with Sharpe, >0.62 in every window), scoring the regime you fear (bear-day accuracy, \$320k net, best rule), grading across many time folds (last fold's best repeats only 29%).
- Not predictive: more candidates k , AUC (0.34, negative in 2023), one combined score (\$167k, worse than accuracy alone), a learned selector under drift (rank IC ~ 0), within-month bench score (0.02 corr).
- Caveat: correlation is not money. Nothing beats never switching strat (\$344k).

Discussion

- Try it yourself! Report a correlation matrix for your evals like you would for a portfolio.
- Before you add eval #N, check it adds information and doesn't duplicate one you already have.
- There's lots more you can do to eval the evals within your system!