

Leakage-Controlled Chronological Evaluation of Hybrid Option Pricing Models

Yuyang Zhao*

Moansh University, Melbourne, Australia

**Corresponding author: Yuyang Zhao.*

Abstract

Option-pricing comparisons can be biased when quote-derived inputs or later outcomes enter model selection. We compare analytical, learned, fixed-blend, and conditional-blend models in a chronology that ends tuning in 2019 and fixes the 2024 to 2025 confirmation before its outcomes are examined. SPY, QQQ, and IWM are assessed in four prespecified sampling streams using raw mean-daily normalized-midpoint MAE. Every model uses features available at prediction time and is scored on the same contract-date keys within each asset-stream cell. The MAE-aligned mixture improves on the validation-optimal fixed blend in 1/4 SPY, 4/4 QQQ, and 4/4 IWM streams; SPY also has reversals and unresolved intervals. Across all 12 asset-stream cells, a simpler winner-probability soft gate has lower macro-average error, while component removals show no stable direction. Date-cluster intervals and a moving-block sensitivity check retain that uncertainty and report unfavorable streams alongside favorable ones. Conditional blending is thus asset- and sample-dependent, not uniformly preferable. The study contributes a leakage-controlled evaluation protocol and a bounded empirical comparison.

Keywords

option pricing, chronological evaluation, temporal leakage, conditional blending, Black–Scholes–Merton

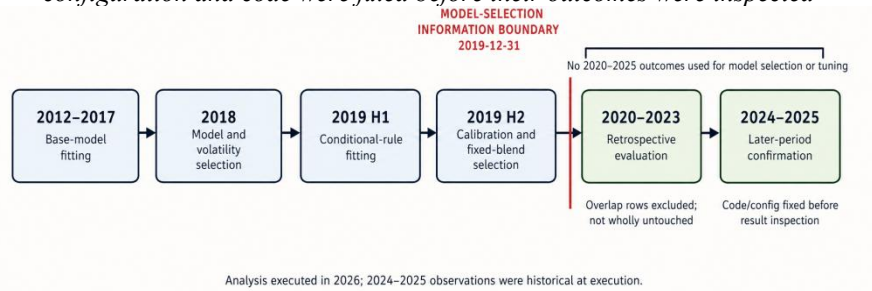
1. Introduction

Analytical and learned option-pricing models draw on different sources of information. Black–Scholes–Merton (BSM) provides a transparent benchmark based on spot, strike, maturity, rates, dividends, and volatility [1, 2]; learned models can fit empirical departures from that benchmark [3, 4]. Their ordering can still depend on the available information and on when each decision is made. Same-row implied volatility or quote fields can reveal the target, random splits can mix market regimes, and final-period outcomes can quietly steer model selection.

Conditional blending makes those timing choices more consequential. A fixed blend uses one validation-selected weight; a conditional gate changes the analytical–learned weight with contract and market features. Such a gate may help when the lower-error expert differs across observations, yet it can also overfit the period used to fit it. The question here is whether conditional blending improves reliably on a validation-optimal fixed blend once experts, features, losses, and selection rules have all been fixed.

We evaluate end-of-day SPY, QQQ, and IWM chains using only variables available at prediction time. Base models are fitted through 2017 and selected in 2018; conditional rules use separate 2019 fitting and calibration blocks. After overlap exclusion, 2020–2023 provides a retrospective held-out evaluation, while the primary evidence comes from a 2024–2025 confirmation fixed before outcome inspection. Figure 1 marks the information boundary. Within that design, the paper gives a timestamp-safe protocol, a matched comparison of analytical, learned, fixed-blend, and conditional-blend models under date-equal MAE, and a three-asset, four-stream test of ranking stability. The evidence remains limited: conditional blending helps consistently for QQQ and IWM, but not SPY, and no gate is claimed to be universally best.

Figure 1: Chronological protocol. The 31 December 2019 marker ends model selection and tuning. The 2024–2025 configuration and code were fixed before their outcomes were inspected



The freeze concerns outcome inspection rather than calendar time. The quotes already existed in 2026, but stream configurations, sampling strata, implementation, and reporting rules were fixed before errors or rankings were read. It is therefore an internal outcome-inspection freeze, not external preregistration.

2. Related Work

Data-driven option pricing includes learning networks, neural forecasting, homogeneity constraints, support-vector regressors, and machine-learning comparisons [3]–[7]. Reviews show that targets, information sets, splits, and analytical comparators differ markedly across these studies [8]; structured networks can also encode financial restrictions [9]. Our question is narrower: do rankings remain stable when contemporaneous quote-derived inputs are excluded and every choice is made before evaluation? Here, BSM serves as an analytical comparator rather than an exact valuation model for end-of-day ETF options.

Conditional blending builds on mixture-of-experts allocation [10, 11]. Disagreement within an ensemble measures predictive dispersion [12], but a score for the winning expert is not necessarily a calibrated probability [13]. We therefore distinguish a winner-probability gate from a mixture weight fitted to smooth MAE. Random forests and CatBoost offer non-neural comparators [14, 15], implemented with established software [16]. Financial evaluation likewise depends on explicit timing and out-of-period tests [17]. We treat each gate as a matched alternative that must hold up across streams.

3. Method

3.1 Prediction Target and Information Boundary

For contract i on date t , let $B_{i,t}$, $A_{i,t}$, and S_t denote the valid bid, ask, and underlying close. The prediction target is

$$y_{i,t} = \frac{B_{i,t} + A_{i,t}}{2S_t}. \quad (1)$$

This normalization puts errors from different dates and assets on a common scale. Predictors are log-moneyness, maturity, option type, selected lagged realized volatility, a lagged maturity-matched rate, and a lagged dividend yield. Same-row implied volatility, quote fields, spread, Greeks, marks, and target proxies are excluded. Rates and volatility are shifted so that source observations precede prediction, and preprocessing and selection are confined to the assigned historical block. Evaluation targets are joined only after frozen predictions have been written for identical contract-date keys.

The analytical anchor is dividend-adjusted BSM [1, 2]. Rates come from the latest pre-quote Treasury curve, interpolated across its 30-, 90-, 180-, 365-, and 730-day maturities [18]; dividend yield uses earlier cash distributions and the preceding close. In the 2018 block, we select among 21-, 63-, and 126-day realized-volatility windows and an exponentially weighted candidate. Because the data use end-of-day joins and the options may permit American exercise, BSM is a benchmark rather than structural ground truth.

3.2 Experts and Mixture Rules

The learned expert is a five-member residual neural ensemble. Each member predicts a standardized BSM residual, which is returned to the original scale and added to the anchor. Candidate hidden layers are $\{32\}$, $\{64,32\}$, and $\{128,64\}$; models use learning rate 10^{-3} , batches of 256, at most 100 epochs, and patience 10. Capacity and stopping epoch are selected in 2018, after which the ensemble is refitted through 2018. Random forest and CatBoost are direct comparators selected in the same block.

All blend rules share the following orientation:

$$\hat{y}_{i,t}(w) = (1 - w_{i,t})p_{i,t}^{\text{BS}} + w_{i,t}p_{i,t}^{\text{ML}}, \quad 0 \leq w_{i,t} \leq 1, \quad (2)$$

In the convex blend, each expert output is normalized; the superscripts distinguish the BSM analytical expert from the learned residual ensemble. The fixed comparator scans 1,001 weights on 2019H2 and selects the lowest date-equal MAE. The winner-probability gate fits a 2019H1 logistic classifier labelled by the lower-error expert. It is calibrated in 2019H2 by Brier score, using pricing error to break ties; the resulting winner probability supplies the learned-expert weight in (2).

The MAE-aligned gate takes standardized contract features, training-support density, and ensemble disagreement as inputs to a sigmoid weight. Support is calculated from 2012–2018 rows, and disagreement is the across-member standard deviation. With n gate-fitting contracts, coefficients β minimize

$$\mathcal{L}(\beta) = \frac{1}{n} \sum_{i=1}^n \sqrt{[y_i - \hat{y}_i(w_i)]^2 + 10^{-6}} + \lambda \|\beta\|_2^2, \quad (3)$$

where n is the number of gate-fitting contracts and the intercept is unpenalized. The 2019H2 grid searches $\lambda \in \{0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$ by date-equal MAE. This weight is not interpreted as a probability. Component removals examine, rather than presume, whether support density and disagreement add value.

3.3 Chronological Protocol

Figure 1 allocates fitting to 2012–2017, base-model selection to 2018, gate fitting to 2019H1, and regularization, calibration, and fixed-blend selection to 2019H2. Models, scalars, features, losses, grids, clipping rules, and metrics are fixed at the end of 2019. The 2020–2023 evaluation omits registered overlapping contract-date rows. Removing them does not remove period-level development knowledge, so this block is described as a retrospective held-out evaluation. For the 2024–2025 confirmation, configurations were fixed before errors or rankings from that period were inspected. This is an internal outcome-inspection freeze rather than external preregistration.

4. Experimental Setup and Results

4.1 Data, Sampling, and Inference

We use end-of-day SPY, QQQ, and IWM chains and underlying histories from the LambdaClass Options Portfolio Backtester release [19]. The raw source files are not redistributed with this submission because the upstream release does not clearly state redistribution terms. The source contains 2024–2025 data, whereas earlier prediction outputs end on 29 December 2023.

The primary filter requires $bid > 0$, $ask > bid$, 7–730 days to expiration, and finite positive spot; it has no moneyness, volume, open-interest, or maximum-spread cutoff. We remove exact duplicate rows and exclude any conflicting quote for a contract-date key. For each asset and calendar year, remaining contracts are stratified by month, option type, log-moneyness, and days-to-expiration bins. Proportional largest-

remainder allocation samples without replacement and retains at most 8,000 observations per year; a shortfall is filled only from remaining eligible contracts in that year. Each asset-stream cell therefore contains 16,000 contracts, yielding 192,000 scored rows across 12 cells. The sampler records eligible and selected stratum counts, allocation shortfalls, and composition summaries. We audit time boundaries, annual caps, contract-date uniqueness, and identical prediction coverage. All tuning rows end by 31 December 2019.

In the definition below, where T denotes the retained trading-date index set and I_t denotes the contracts on date t , the primary estimand is raw date-equal normalized-midpoint MAE:

$$\text{MAE}_{\text{day}} = \frac{1}{|T|} \sum_{t \in T} \frac{1}{|I_t|} \sum_{i \in I_t} |y_{i,t} - \hat{y}_{i,t}|. \quad (4)$$

Paired date-cluster intervals use 5,000 resamples, with a 20-day moving-block bootstrap as the dependence sensitivity check. Uniform clipping is secondary to the raw ranking.

4.2 Pre-outcome-inspection Frozen Main Stream

Table 1 gives the S0 confirmatory results. CatBoost leads SPY, the MAE-aligned mixture leads QQQ, and the soft gate leads IWM, so no method is a single headline winner.

Table 1: S0 frozen raw mean-daily normalized-midpoint MAE ($\times 10^{-3}$). Lower is better.

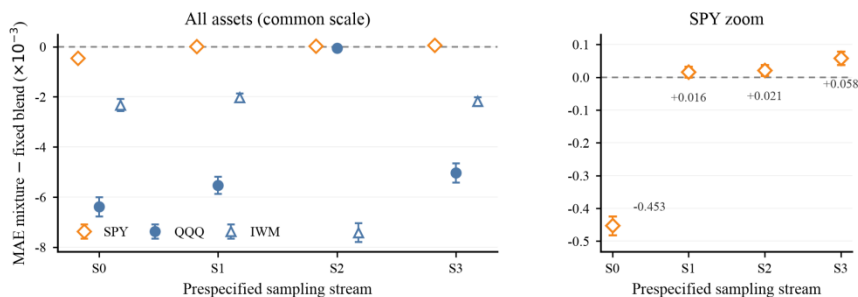
Method	MAE ($\times 10^{-3}$)		
	SPY	QQQ	IWM
BSM, selected lagged RV	8.333	8.237	7.147
Random forest	6.240	8.481	7.877
CatBoost	5.897	8.347	6.483
Residual ensemble	8.122	14.351	8.121
Validation-optimal constant blend	8.121	12.376	8.121
Winner-probability soft gate	7.199	6.452	5.743
MAE-aligned conditional mixture	7.669	5.994	5.802

In S0, the MAE-aligned mixture is 0.453, 6.383, and 2.318 below the fixed blend for SPY, QQQ, and IWM, respectively. The paired date-cluster 95% intervals lie 0.424–0.481, 6.001–6.775, and 2.084–2.553 below zero. The moving-block sensitivity check points in the same direction for this pre-outcome-inspection stream, but neither comparison establishes leadership over the broader model set.

4.3 Sampling-stream Stability

Figure 2 summarizes stability across streams. The MAE-aligned mixture beats the fixed blend in 1/4 SPY, 4/4 QQQ, and 4/4 IWM cells. Date-cluster intervals label eight cells as improvements, two SPY cells as worsenings, and two as unresolved; MBB-20 identifies eight improvements and four unresolved cells. SPY is 0.453 below the fixed blend in S0, then 0.016, 0.021, and 0.058 above it in S1–S3. QQQ remains below in every stream, although S2 is unresolved, and IWM improves in every stream.

Figure 2: Conditional-minus-fixed MAE across frozen sampling streams. Negative values favour the MAE-aligned mixture; SPY reverses after S0, whereas QQQ and IWM remain negative



The 2020–2023 retrospective evaluation has the same 1/4, 4/4, and 4/4 pattern. Across the 12 confirmation cells, the soft gate has lower macro-average MAE (0.006550) than the MAE-aligned mixture (0.006620). Removing support density lowers error in no SPY stream and in half of QQQ/IWM streams; removing

disagreement lowers error in three SPY and half of QQQ/IWM streams. The ablation is therefore diagnostic, not evidence that either reliability proxy contributes positively for every asset or sample.

5. Discussion

The results support an evaluation conclusion, not a universal claim about model quality. The MAE-aligned mixture is consistently better than the fixed blend for QQQ and IWM, but not for SPY; the soft gate has the lower 12-cell macro average. The later confirmation reproduces the retrospective asset pattern under a sampler fixed before outcome inspection, yet its four streams are deterministic alternative samples from one quote source rather than independent markets. Reporting only S0 would hide the SPY reversals, while reporting only the MAE-aligned row would hide the soft gate's aggregate advantage.

The full ranking illustrates the distinction: CatBoost leads S0 SPY, the MAE-aligned mixture leads S0 QQQ, and the soft gate leads S0 IWM. Optimizing the mixture directly against smooth absolute error does not guarantee the lowest aggregate error. The component check is not a mechanism result either. Support density has no stable positive contribution across assets, whereas ensemble disagreement is often neutral or adverse. Conditional weights may help in particular asset-sample combinations, but neither their functional form nor these diagnostic inputs should be treated as inherently reliable.

The protocol retains adverse evidence instead of screening it away. Had only S0 been reported, every date-cluster interval would favour the MAE-aligned mixture over the fixed comparator. The additional streams show two SPY worsenings and one unresolved SPY comparison. By prespecifying assets, streams, raw predictions, and interval rules, we keep those outcomes in the comparison rather than treating them as post hoc exclusions.

Date-equal aggregation stops contract-rich dates from dominating the estimand, and paired resampling retains within-date comparisons. The moving-block check tests whether short-range temporal dependence changes interval classification without replacing the primary date-cluster analysis. The raw mirror is hash-verified within this study, yet upstream provenance and redistribution terms remain incomplete. We therefore present the release as reproducible processing evidence rather than independent verification of the original market quotes. End-of-day joins leave timestamp ambiguity, and ETF options may permit American exercise, so BSM remains an analytical benchmark. The 2024–2025 block evaluates temporal and sampling stability, not independent-source generalization; four deterministic resamples of one source cannot become four independent market studies.

6. Conclusion

With a leakage-controlled chronology, the MAE-aligned mixture improves on the fixed blend in every QQQ and IWM confirmatory stream, but in only one SPY stream. The simpler soft gate has lower 12-cell macro-average error, and component effects are not stable. Conditional blending is asset- and sample-dependent, rather than uniformly preferable. The protocol provides a reproducible basis for subsequent comparisons using independently sourced assets and pre-specified evaluation periods.

References

- [1] F. Black and M. Scholes, "The pricing of options and corporate liabilities," *Journal of Political Economy*, vol. 81, no. 3, pp. 637–654, 1973. doi: [10.1086/260062](https://doi.org/10.1086/260062)
- [2] R. C. Merton, "Theory of rational option pricing," *The Bell Journal of Economics and Management Science*, vol. 4, no. 1, pp. 141–183, 1973. doi: [10.2307/3003143](https://doi.org/10.2307/3003143)
- [3] J. M. Hutchinson, A. W. Lo, and T. Poggio, "A nonparametric approach to pricing and hedging derivative securities via learning networks," *The Journal of Finance*, vol. 49, no. 3, pp. 851–889, 1994. doi: [10.1111/j.1540-6261.1994.tb00081.x](https://doi.org/10.1111/j.1540-6261.1994.tb00081.x)
- [4] J. Yao, Y. Li, and C. L. Tan, "Option price forecasting using neural networks," *Omega*, vol. 28, no. 4, pp. 455–466, 2000. doi: [10.1016/S0305-0483\(99\)00066-3](https://doi.org/10.1016/S0305-0483(99)00066-3)

- [5] R. Garcia and R. Gençay, "Pricing and hedging derivative securities with neural networks and a homogeneity hint," *Journal of Econometrics*, vol. 94, no. 1–2, pp. 93–115, 2000. doi: [10.1016/S0304-4076\(99\)00018-4](https://doi.org/10.1016/S0304-4076(99)00018-4)
- [6] X. Liang, H. Zhang, J. Xiao, and Y. Chen, "Improving option price forecasts with neural networks and support vector regressions," *Neurocomputing*, vol. 72, no. 13–15, pp. 3055–3065, 2009. doi: [10.1016/j.neucom.2009.03.015](https://doi.org/10.1016/j.neucom.2009.03.015)
- [7] C.-F. Ivaşcu, "Option pricing using machine learning," *Expert Systems with Applications*, vol. 163, p. 113799, 2021. doi: [10.1016/j.eswa.2020.113799](https://doi.org/10.1016/j.eswa.2020.113799)
- [8] J. Ruf and W. Wang, "Neural networks for option pricing and hedging: A literature review," *The Journal of Computational Finance*, vol. 24, no. 1, pp. 1–46, 2020. doi: [10.21314/JCF.2020.390](https://doi.org/10.21314/JCF.2020.390)
- [9] Y. Yang, Y. Zheng, and T. M. Hospedales, "Gated neural networks for option pricing: Rationality by design," in *Proceedings of the thirty-first AAAI conference on artificial intelligence*, vol. 31, pp. 52–58, 2017. doi: [10.1609/aaai.v31i1.10505](https://doi.org/10.1609/aaai.v31i1.10505)
- [10] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991. doi: [10.1162/neco.1991.3.1.79](https://doi.org/10.1162/neco.1991.3.1.79)
- [11] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, vol. 6, no. 2, pp. 181–214, 1994. doi: [10.1162/neco.1994.6.2.181](https://doi.org/10.1162/neco.1994.6.2.181)
- [12] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in neural information processing systems*, vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html. Accessed: July 30, 2026
- [13] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th international conference on machine learning*, vol. 70, pp. 1321–1330, PMLR, 2017. [Online]. Available: <https://proceedings.mlr.press/v70/guo17a.html>. Accessed: July 30, 2026
- [14] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. doi: [10.1214/aos/1013203451](https://doi.org/10.1214/aos/1013203451)
- [15] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Advances in neural information processing systems*, vol. 31, Curran Associates, Inc., 2018. [Online]. Available: <https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html>. Accessed: July 30, 2026
- [16] F. Pedregosa et al., "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011. [Online]. Available: <https://www.jmlr.org/papers/v12/pedregosa11a.html>. Accessed: July 30, 2026
- [17] J. Ruf and W. Wang, "Hedging with linear regressions and neural networks," *Journal of Business & Economic Statistics*, vol. 40, no. 4, pp. 1442–1454, 2022. doi: [10.1080/07350015.2021.1931241](https://doi.org/10.1080/07350015.2021.1931241)
- [18] U.S. Department of the Treasury, "Daily treasury par yield curve rates," Online data and methodology, 2026. [Online]. Available: https://home.treasury.gov/resource-center/data-chart-center/interest-rates/TextView?type=daily_treasury_yield_curve. Accessed: July 29, 2026
- [19] LambdaClass, "Options portfolio backtester: Data release v1," GitHub release and source repository, 2026. [Online]. Available: https://github.com/lambdaclass/options_portfolio_backtester/releases/tag/data-v1. Accessed: July 29, 2026

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access

This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

