Supplementary Methods: Experimental Evidence and Aggregate Analyses

Analysis date: 2026-09-18. The supplementary analyses use experiment_evidence.json, with values extracted from completed run manifests and saved predictions. No additional training or inference was performed. The aggregate CSV files contain date-level model or portfolio statistics, without symbol-level rows or news text.

Scope of the Experimental Evidence

The observed evidence consists of N0 seed 42 across three expanding-time folds and four model variants on Fold 0 at seed 42. The configured seeds 314159 and 271828 are planned settings rather than completed experiments. All reported validation scores were involved in selecting the best checkpoint. The run manifests contain the flag pilot_only because the complete planned model/seed/fold matrix was not executed. They also record zero development-holdout inference passes and no access to sealed rows. These records represent development-validation evidence rather than an independent test, untouched out-of-sample performance, prospective trading performance, or a complete multi-seed benchmark.

Fold	Training interval	Actual training rows / dates	Validation interval	Validation rows / dates	Best step	Final step
0	2010-03-01 to 2011-12-09	19,512 / 306	2011-12-27 to 2013-10-17	65,036 / 448	1,715	1,715
1	2010-03-01 to 2013-10-17	89,297 / 775	2013-11-01 to 2015-08-24	107,457 / 455	3,000	4,500
2	2010-03-01 to 2015-08-24	201,797 / 1,240	2015-09-09 to 2017-07-10	126,276 / 447	7,500	9,003

Fold 0 initially routed 19,847 training rows across 318 dates. The minimum of 30 symbols per training date excludes 335 rows across 12 dates, leaving the 19,512 rows used for training and scaler fitting. Training observations in later folds are cumulative and cannot be summed as a count of unique observations. The 298,769 validation rows and 1,350 validation dates cover disjoint validation intervals. Scalers were fitted to the training split of each fold.

Numeric Model and Optimization

The public N0 configuration records 23 numeric/context features, a missingness indicator for each, and two transformed count inputs, yielding 48 inputs in total. The network dimensions are 48 → 64 → 32 → 1, with LayerNorm and GELU in the numeric tower, dropout 0.1, and 5,345 parameters. The transformations of count inputs and the exact preprocessing order are defined in inference.py. The scalar output is a learned ranking signal rather than a percentage return.

Configured supervised training uses a main loss comprising 0.25 date-level Huber loss and 0.75 RankNet loss, Huber beta 1, RankNet temperature 1, at most 4,096 deterministic rank-gap-stratified pairs per date, equal date weighting, a maximum of eight epochs, validation patience four, minimum IC improvement 0.0001, head learning rate 0.0001, weight decay 0.01, linear warmup/decay with 6% warmup, gradient norm limit 1, and validation every 750 optimizer updates plus epoch-end checks. The extracted

hyperparameters object contains the complete shared configuration. The executed N0 date loss is 1.10 times the stated 0.25/0.75 main-loss combination. Update-group losses use the fixed divisor 3 established by the completed training date schedules; the final group of an epoch may contain fewer dates. Text-specific settings such as LoRA and auxiliary text loss apply to the text/fusion branches rather than to N0 trainable components.

N0 training wall times recorded in the three fold manifests are 273.983, 823.876, and 1,598.322 seconds, respectively. These run-specific measurements do not establish hardware-independent execution speed. Fold 0 text/fusion comparison wall times are recorded in the JSON. GPU peak-allocation records in the fold manifests are implementation measurements rather than estimates of total device memory or deployment requirements.

Checkpoint Selection and Learning Curves

training_validation_curves.csv records the optimizer step, zero-based epoch, validation IC, check reason, and whether the checkpoint was selected. It contains all four Fold 0 variants and N0 Folds 1–2. These are checkpoint-selection validation curves rather than untouched test trajectories. N0 Fold 0 improves from 0.001822 at step 216 to 0.016330 at step 1,715; Folds 1 and 2 select steps 3,000 and 7,500 before stopping later. For all three selected N0 checkpoints, the maximum absolute prediction difference after saving and reloading is zero in the original inference environment.

Post-Hoc Moving-Block Uncertainty Analysis

The post-hoc uncertainty analysis uses daily IC vectors saved in validation_metrics_after_reload.daily. N0 daily IC was independently recomputed from the saved validation predictions and target ranks, matching the recorded vectors within 1e-12. Fold 0 model comparisons were checked against the same 65,036 row keys, dates, and targets. Daily IC differences are calculated on exactly matching dates before resampling.

For block length L in {5, 10, 20}, the algorithm forms overlapping non-circular blocks of consecutive observed eligible evaluation dates. It samples ceil(T/L) blocks with replacement, concatenates them, truncates the result to T dates, and computes the sample mean. This procedure is repeated 2,000 times using NumPy default_rng and the deterministic seeds recorded in the JSON. The interval endpoints are the 2.5th and 97.5th percentiles. The combined N0 interval resamples separately within each fold and preserves each fold's date count. Blocks do not cross fold boundaries. Missing evaluation sessions are not imputed, so L denotes observed eligible dates rather than assuming that every exchange session is present.

For the 10-date block specification:

Quantity	Observed mean	Pointwise 95% percentile interval
N0 combined mean daily IC	0.020877	[0.013129, 0.028403]
Fold-0 T1 minus N0 daily IC	-0.012366	[-0.026121, 0.002959]
Fold-0 F0 minus N0 daily IC	-0.000461	[-0.006870, 0.006988]
Fold-0 zero-text minus N0 daily IC	0	[0, 0]

The paired-difference intervals for T1 and F0 include zero at all three block lengths. The evidence supports the descriptive statement that this experiment observed no improvement over N0 while retaining uncertainty about the true difference. It does not establish statistical superiority of N0, equivalence

between F0 and N0, or the absence of predictive information in news text. The zero-text control is exact by construction, with verified equality of outputs; its degenerate interval is not an independent empirical replication.

These intervals are conditional on already-selected checkpoints, features, sample screens, and prior development. They do not correct checkpoint-selection bias, model-search bias, reuse of development data, or cross-fold dependence arising from overlapping training histories. The classical IC t-statistics in earlier summaries are consequently treated as descriptive statistics rather than unqualified evidence of statistical significance.

Post-Hoc Calendar-Year Summaries

new_posthoc_annual_ic groups the same N0 selected-checkpoint validation observations by calendar year without fitting a new model. Some calendar years combine predictions from two selected fold checkpoints. Mean IC is 0.020680 (2012, 244 dates), 0.013904 (2013, 242), 0.033828 (2014, 252), 0.017405 (2015, 227), 0.011128 (2016, 252), and 0.031386 (2017 through July 10, 130). The 2011 tail contains only three dates and constitutes explicitly partial coverage rather than a complete annual estimate. These are descriptive aggregates, not separate independent yearly trials or a complete calendar-day strategy.

Portfolio Diagnostics and Construction Definitions

The original validation_long_short_daily.parquet and validation_long_short_summary.json use top/bottom 20% quintiles. Their combined gross arithmetic annualized return is 17.0543%, and their Sharpe ratio is 1.1431. The public summary's 26.7251% and 1.3355 instead refer to top/bottom 10% deciles selected from the exploratory sweep over side fractions of 5%, 10%, 15%, 20%, and 30%. The quintile and decile values refer to distinct portfolio constructions.

Both the 10% and 20% daily series were reconstructed from existing validation predictions and existing raw open-to-close return targets, using the original deterministic ordering (prediction, vendor_symbol_day_id), max(1, floor(universe_size * fraction)) assets per side, and equal within-side weights. Only date-level aggregate returns are exported. The reconstructed arithmetic means, annualized volatility, Sharpe ratios, and positive-day shares match the original sweep to 1e-12 at 0, 1, 2, and 5 bp per side.

The convention uses long notional exposure of 1, absolute short notional exposure of 1, and gross exposure of 2. Opening and closing both sides produces a fixed daily deduction of 4 * c / 10000, where c is the one-way cost in basis points. One bp therefore deducts 0.0004 = 0.04% of the specified capital basis per observed trading day, rather than 0.01%. Reported annualized arithmetic return is 252 times the observed-day mean, and annualized volatility is the sample standard deviation multiplied by sqrt(252). The Sharpe ratio uses a zero risk-free benchmark under the same convention.

For the exploratory 10% construction, annualized returns are 26.7251%, 16.6451%, 6.5651%, and -23.6749% at 0, 1, 2, and 5 bp per side, respectively. The corresponding Sharpe ratios are 1.3355, 0.8318, 0.3281, and -1.1831. The estimated mean-return break-even cost is 2.6513 bp per side. The 10% fraction achieved the highest gross Sharpe ratio among the five examined fractions, while the 5% fraction had a higher arithmetic return. The decile result therefore remains a development-selected portfolio diagnostic.

The supplementary wealth and drawdown summaries concatenate observed validation dates under this return convention, omit gaps, and apply the cost deductions described above. They are illustrative portfolio

paths. They exclude market impact, stock-borrow expenses and availability, fill uncertainty, and capacity effects, and therefore do not establish implementable realized profit.

Evidence Files and JSON Schema

- experiment_evidence.json: master values, exact configuration, sample windows, all fold/core4 records, uncertainty intervals, calendar-year summaries, the original portfolio sweep, the reconstructed decile summary, source hashes, and verification checks.
- daily_ic_aggregate.csv: fold, model_id, decision_date, firm_days, spearman; N0 across all folds and the other three models on Fold 0.
- portfolio_daily_aggregate.csv: fold, decision_date, side_fraction, universe_size, side_size, long_short_gross_return; fractions 0.1 and 0.2 only.
- training_validation_curves.csv: fold, model_id, epoch_zero_based, optimizer_step, mean_daily_spearman_ic, reason, selected_as_best.
- extract_experiment_evidence.py: reproducible read-only analysis requiring NumPy, pandas, PyArrow, and PyYAML; reads the original local artifacts and writes only aggregate outputs.

Master JSON keys: model_config, published_evaluation, original_validation_summary, hyperparameters, actual_seed, planned_seeds_are_not_executed_evidence, dataset_contract, primary_screen, run_status, n0_folds, fold0_core4, training_validation_curves, paired_fold0_uncertainty, new_posthoc_annual_ic, n0_mean_ic_uncertainty, bootstrap_method, matched_zero_text_control, original_portfolio_sweep, original_quintile_portfolio_summary, new_posthoc_decile_curve_summary, integrity_checks, portfolio_interpretation, aggregate_artifacts, source_hashes, analysis_runtime.

Each interval object contains block_length_eligible_dates, replicates, seed, observed_mean, percentile_95_ci_low, percentile_95_ci_high, and bootstrap_standard_error. Source SHA-256 values refer to the original files as of the analysis date. Fold-level values in the master JSON mirror the recorded training runs. The extraction script reads original run manifests for model and learning-curve values; running it requires those original project artifacts, which are not part of the aggregate supplement.

