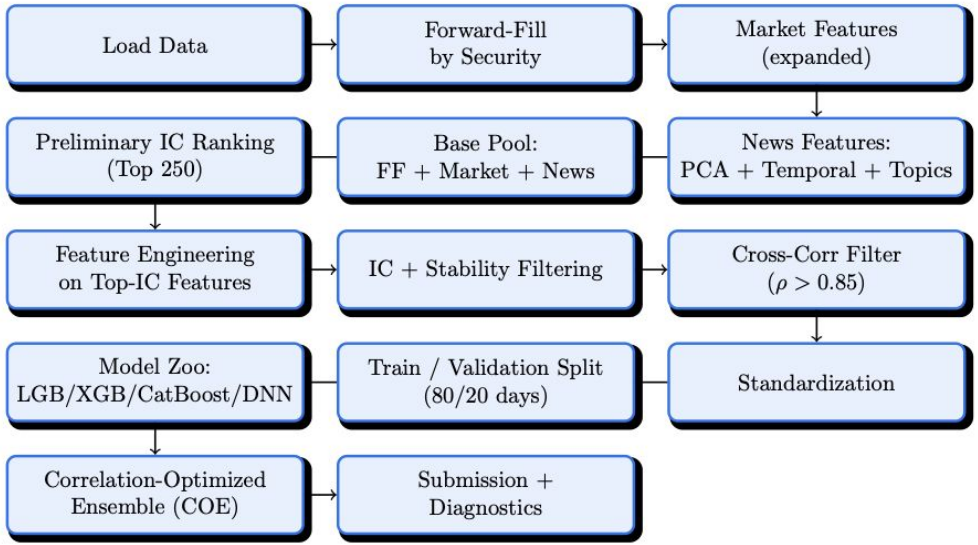
Final Presentation

ThinkSystemSR780aV3

Michael Cao, Kexin Deng, Yichen Gao

Project Pipeline and Feature Engineering

400 Alpha Signals ⇨ Forward-Fill Grouped By Stock IDs ⇨ Market Feature Engineering (55 Features)



Category	Examples	Count
Liquidity	vol/avg_vol, log(volume), sqrt(volume)	5
News Flow	news_volume, sentiment, vol×sentiment	4
Announcement Timing	days_since, days_until, 1/(days+1), log transforms	6
Short-Horizon Returns	rret, rret–SPY, sign(rret), abs(rret)	4
Multi-Horizon Returns	ret2, ret5, SPY-adjusted, sign/abs, VIX interactions	10+
Return Interactions	rret×ret2, ret2×ret5, momentum×market	5
Macro/VIX Regime	VIX, VIX×return, log(VIX), sqrt(VIX), normalized VIX	5
Sector/Industry Encoding	raw IDs + sin/cos positional encodings	4
Market-Relative Signals	returns normalized by VIX	4

Project Pipeline and Feature Engineering

News Feature Engineering (342 Features)

Component	Description	Count
PCA Embeddings	768 → 64 dimensions	64
Temporal Features	1d/3d diff, rolling mean/std (window=5), cosine similarity to previous day + window mean	258
Topic Clusters	K-Means (20 clusters) → one-hot	20
Total	—	**342**

400 alphas + 55 market features + 342 news features

= 797 total

⟹ IC filtering & nonlinear expansion

IC Ranking & Nonlinear Feature Expansion

Preliminary IC Ranking (Top 250)

- Compute full-sample IC for each feature
- Select top-|IC| 250

Nonlinear Feature Expansion (~2000 Features)

- Included Expansions:

Type	Description
Polynomial transforms	squared, abs, sign-sqrt, clipping
Feature interactions	top 50 × (VIX, rret, SPY)
Risk-sensitive nonlinearities	return × volatility, alpha × market

Stability Filters: Final 237 Features

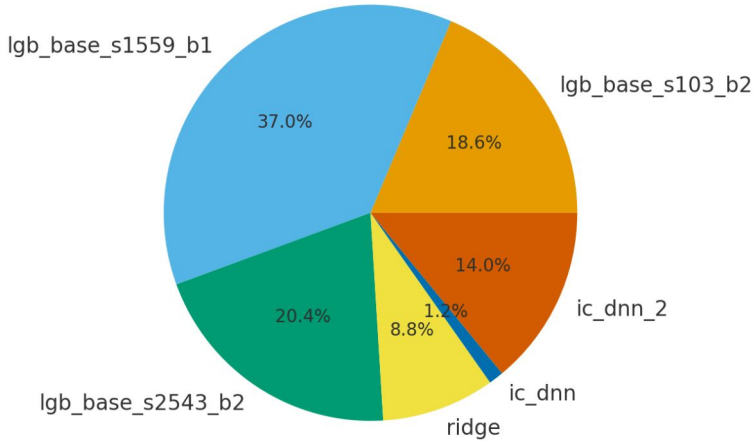
1. IC Stability Filter
   - Split time into 5 blocks
   - Keep features with consistent IC sign across ≥3 blocks
   - Ensures temporal robustness

2. Cross-Correlation Filter
   - Compute feature–feature correlation matrix
   - Group correlated features
   - Keep the group leader with highest |corr(y)|

Model Zoo Construction

Key Techniques: Random feature bagging, Multiple seeds, CV via GroupKFold
Ensemble: Correlation-Optimized Ensemble (COE) with Weight Caps

LightGBM	6 seeds x 3 bags
XGBoost	6 seeds x 3 bags
Linear	Ridge, Lasso
Deep Learning	IC-loss DNN
Others	CatBoost, TabNet



Tested a wide range of feature configurations (60 → 2,000+) before finalizing a stable filtered set.
Evaluated automated feature tools (AutoFeat, MI ranking, SULOV/MRMR), but results were noisy or unstable.
Explored latent factor methods (PCA/ICA, SGM, Sparse Autoencoder), ultimately removed due to weak OOS performance.
Tried multiple ML/DL models (ExtraTrees, RF, MLP, CatBoost, Lasso, TabNet) that underperformed.
Compared ensemble approaches (NNLS, correlation-weighted ranks, PCA-decorrelated softmax Ridge) before selecting COE with weight control.