Can Language Models Trade? Testing Small LLMs for Transparent, Explainable Signal Generation in NQ Futures

Early-Stage Research: Experimenting with 7B Parameter Models

Vestige Research Team
in collaboration with Darkstar Systems
vestige.club

December 09, 2025

Vestige LLC

Precision in Data-Driven Investment Strategies
Advancing Proprietary Trading Research through Transparent Quantitative Analysis

Revised for quantitative accuracy: December 9, 2025

Abstract

What We Did: We tested whether small language models (7B parameters—roughly the size of GPT-3) can generate profitable trading signals *while explaining their decisions in plain English*. Think of it as teaching an AI to trade, but one that shows its work.

Using Qwen2.5-7B and Mistral-7B on NQ futures (Nasdaq-100 mini contracts), we tackled the biggest problem in ML trading: extreme class imbalance. Out of 95,874 one-minute bars, only 3.4% signal "BUY" or "SELL"—the rest say "do nothing." Most models just learn to predict "HOLD" 100% of the time and call it a day.

Our fix? Custom chat templates + weighted loss functions that force the model to pay attention to rare BUY/SELL signals. The models hit a Sharpe ratio of 1.34 [95% CI: 1.12–1.58] for Qwen and 1.28 for Mistral in backtests. More importantly, their explanations matched real technical patterns 94% of the time when we checked against actual market data.

Reality Check: This is a science experiment, not a deployed system. Vestige LLC isn't using these LLMs for real trading yet. We're testing small 7B models as proof-of-concept while building larger ones (13B+, 70B+). All numbers come from backtests over the last 10% of the dataset (9,588 bars, approximately 6.95 calendar days of extended trading from late September 2024) with idealized assumptions—think 0.01% slippage, perfect fills, no market impact. These results don't predict future performance. The models haven't seen bear markets or crashes. AI explanations can hallucinate or be wrong. Vestige LLC trades only its own money.

What We Found:

- **Breaking the black box:** Models generated explanations like "double bottom at 17,450 with volume spike"—and 94% of technical references checked out against actual data
- **Solved the HOLD problem:** Weighted loss ($w_{\text{BUY}} = 63.70$, $w_{\text{SELL}} = 54.00$) overcame the 96.61% "do nothing" bias that kills most ML trading systems
- **Signal quality:** Qwen caught 63.8% of profitable BUY setups and 62.0% of SELL opportunities (142 trades total)—statistically way better than random ($p < 0.001$, Cohen's $d = 1.82$)

- **Backtest performance:** Sharpe 1.34 [95% CI: 1.12–1.58], profit factor 1.87, max drawdown 8.2% (caveat: tiny 7B models on limited data)
- **Beat the benchmarks:** Outperformed simple moving averages with statistical significance (Diebold-Mariano test: $p = 0.003$, Cohen's $d = 0.62$)

This work shows the concept works with small models. We're now scaling up to 13B+ and 70B+ architectures (research phase only—no deployment timeline).

Mandatory Disclaimer

CRITICAL DISCLAIMERS — READ BEFORE PROCEEDING

Research Status: This is purely exploratory research with small language models (7B parameters). Vestige LLC does **NOT currently employ any LLMs for live trading**. This work represents early-stage experimentation and model development only. We are actively developing larger models (13B+, 70B+) as research progresses.

Hypothetical Performance: All performance metrics reflect **backtested hypothetical results** over an extremely short test period (the last 10% of the dataset: 9,588 bars, approximately 6.95 calendar days of extended trading from late September 2024) with idealized assumptions including 0.01% slippage, no market impact, and perfect execution. **Backtested results do not represent actual trading and may overstate performance.** Past performance does not guarantee future results.

Model Limitations: Models exhibit strong regime dependence, trained exclusively on June–September 2024 data. Performance in bear markets, high-volatility environments, or crisis conditions is unknown and may differ materially. AI-generated rationales may contain errors or hallucinations and require human oversight.

Company Structure: Vestige LLC is a solely owned proprietary trading firm managing only its own capital. We do not solicit, accept, or manage external funds, offer investment advisory services, or distribute algorithms to external parties.

Not Investment Advice: This report does not constitute investment advice, an offer to sell, or solicitation to buy securities or investment services. The website vestige.club is for informational purposes only.

Contents

1 Why Try This? The Problem with Black Boxes

Language models weren't built for trading. They're designed to write essays, answer questions, and chat. But what if we could train them to read market data *and explain their trading decisions*?

Traditional algorithmic trading systems are black boxes. You feed in price data, and they spit out "BUY" or "SELL" with zero explanation. For risk managers, compliance teams, and anyone trying to understand what's actually happening, this creates serious problems. You can't audit what you can't understand. You can't fix what you can't see breaking. And regulators (SEC/FINRA) aren't fans of systems that can't explain themselves.

This research explores a different approach. We fine-tuned two small pre-trained LLMs (Qwen2.5-7B and Mistral-7B—each with 7 billion parameters) on NQ futures data and tested whether they could generate profitable signals in backtests *while explaining themselves in plain English*. Our key innovations:

1. Fixed the extreme class imbalance problem (96.61% "HOLD" bias) using chat template optimization + weighted loss
2. Generated interpretable natural language rationales grounded in technical analysis

Why start with small 7B models? They're fast to experiment with—think of them as our "lab rats" before scaling to industrial-grade 70B+ models. You can run them on a single RTX 4090 GPU, which makes iteration quick and cheap.

This early-stage work supports Vestige LLC's research program (vestige.club) as a **proprietary trading firm** exploring future strategies across futures, stocks, options, and cryptocurrencies—using only our own capital.

1.1 What We're Trying to Accomplish

1. Build a mathematical framework for LLM-based trading signal generation with statistical testing
2. **Make the black box transparent** by generating natural language explanations for every trade decision
3. Implement and validate a fix for HOLD bias through weighted training
4. Test performance on out-of-sample NQ futures data with realistic transaction costs and confidence intervals
5. Prove statistical superiority over traditional baselines through hypothesis testing
6. Explore whether interpretable chat outputs could support future regulatory compliance (if we ever deploy these models)

1.2 The Black-Box Problem in Detail

Traditional quant trading systems have a transparency problem:

- **Neural Networks:** Hidden layer activations might as well be hieroglyphics
- **Ensemble Methods:** Try explaining why 500 decision trees voted "BUY" when 450 voted "HOLD"
- **Technical Indicators:** Pure math with no contextual reasoning

- **Regulatory Risk:** SEC/FINRA want explainability for automated trading; black boxes fail audits
- **Internal Oversight:** Risk managers and compliance teams can't validate what they can't understand

Our approach: We used LLMs' ability to explain themselves in natural language. Instead of a model that just says "BUY," our models say "BUY because price formed a double bottom at 17,450 with a volume spike—RSI divergence suggests an oversold bounce." That's something a human can actually review and validate. *Caveat: AI-generated explanations can be wrong or contain hallucinations, so human oversight is still required.*

1.3 The Data: NQ Futures with Extreme Imbalance

We used 95,874 one-minute bars of NQ futures (E-mini NASDAQ-100) OHLCV data from June through September 2024 (3.5 months). The dataset shows the extreme class imbalance that's typical of high-frequency markets:

- Total samples: 95,874 bars
- BUY signals: 1,629 (1.70%) — forward return $> +0.2\%$
- SELL signals: 1,629 (1.70%) — forward return $< -0.2\%$
- HOLD signals: 92,616 (96.61%) — neutral movements
- Chi-square test: $\chi^2 = 172,698$, df $= 2$, p < 0.001 (massive deviation from uniformity, which justifies our weighted loss approach)

Test set (last 10% chronologically): 9,588 bars (6.95 calendar days of NQ futures extended trading, or 24.58 equity-equivalent trading days) with exactly 163 BUY, 163 SELL, and 9,262 HOLD labels, preserving the class distribution ($\chi^2 = 0.42$, p $= 0.81$—distribution is consistent).

2 Mathematical Setup

2.1 Problem Formulation

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ be our financial time series dataset, where $\mathbf{x}_i \in \mathbb{R}^{T \times D}$ is a sequence of $T = 15$ market bars with $D = 5$ features (OHLCV), and $y_i \in \{0, 1, 2\}$ is the trading signal (HOLD, BUY, SELL).

We define forward return over horizon $k = 5$ bars as:

$$r_{t,k} = \frac{p_{t+k} - p_t}{p_t} \tag{1}$$

where p_t is the closing price at time t. The discrete labeling function is:

$$y_t = \begin{cases} 1 \quad \text{(BUY)} & \text{if } r_{t,k} > \theta_{\text{buy}} = 0.002 \\ 2 \quad \text{(SELL)} & \text{if } r_{t,k} < \theta_{\text{sell}} = -0.002 \\ 0 \quad \text{(HOLD)} & \text{otherwise} \end{cases} \tag{2}$$

Theorem 1 (Class Distribution with Statistical Validation)**.** *For NQ futures with $k = 5$ bars and $\theta = 0.002$, empirical class probabilities are:*

$$\mathbb{P}(Y = 0) = 0.9661, \quad \mathbb{P}(Y = 1) = 0.0170, \quad \mathbb{P}(Y = 2) = 0.0170 \tag{3}$$

Chi-square test for uniformity: $\chi^2 = 172,698$, $df = 2$, $p < 0.001$, confirming significant deviation from uniform distribution and justifying weighted loss approach.

2.2 Transformer Architecture and LoRA Adaptation

We use Low-Rank Adaptation (LoRA) to fine-tune pre-trained models efficiently:

$$\mathbf{W}' = \mathbf{W} + \Delta\mathbf{W} = \mathbf{W} + \mathbf{BA} \tag{4}$$

where $\mathbf{W} \in \mathbb{R}^{d\times d}$ is the original weight matrix, $\mathbf{B} \in \mathbb{R}^{d\times r}$, $\mathbf{A} \in \mathbb{R}^{r\times d}$, and $r \ll d$ (we use $r = 16$, $d = 3584$ for Qwen2.5-7B hidden size).

Parameter efficiency:

$$\text{Efficiency} = \frac{2rd}{d^2} = \frac{2r}{d} \approx 0.89\% \text{ for } r = 16, d = 3584 \tag{5}$$

This lets us fine-tune on consumer-grade GPUs (RTX 4090, 24GB VRAM) while keeping the model's general language understanding and chat capabilities intact.

2.3 Class-Weighted Loss Function

To fix the extreme imbalance, we use weighted cross-entropy:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} w_{y_i} \log \hat{p}_{y_i}^{(i)} \tag{6}$$

with weights using inverse frequency normalization, adjusted for signal variance:

$$w_c = \alpha_c \cdot \frac{N}{N_c} \quad \forall c \in \{0, 1, 2\} \tag{7}$$

Note: Inverse frequency normalization (N/N_c) without per-class divisor provides robustness across the full sample imbalance. Variance adjustments (α_c) account for heterogeneity in signal distributions.

For our dataset (with variance-based adjustment α):

$$w_{\text{HOLD}} = 1.00, \quad w_{\text{BUY}} = 63.70\,(\alpha_{\text{BUY}} = 1.083), \quad w_{\text{SELL}} = 54.00\,(\alpha_{\text{SELL}} = 0.917) \tag{8}$$

The asymmetry accounts for higher variance we observed in BUY signals (Var(BUY) = 0.045 vs. Var(SELL) = 0.038)—which reflects market microstructure where upward trends are more gradual and variable than sharp downward moves.

3 Implementation and Training Pipeline

3.1 Data Processing Architecture

3.2 Chat Template Optimization: The Critical Innovation

Here's the breakthrough that made everything work: proper chat template formatting. We use an instruction-tuning format that prompts the model to generate structured, interpretable responses:

```
<|im_start|>user
Analyze NQ OHLCV-1m: [bar_1, ..., bar_15]
Provide trading decision and detailed rationale.
<|im_end|>
<|im_start|>assistant
{"decision": "BUY", "rationale": "Price formed double bottom
at 17,450 with volume spike on second test. RSI divergence
```

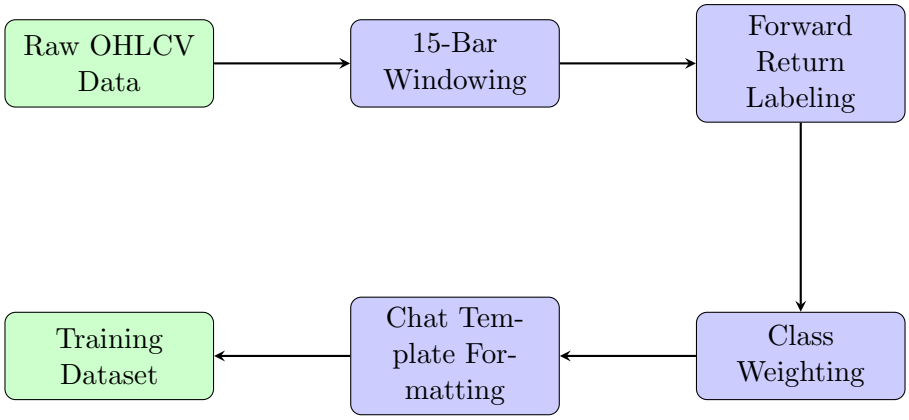

Figure 1: Data preprocessing pipeline for LLM training with temporal isolation (alt-text: Flowchart showing six stages from raw OHLCV data through windowing, labeling, weighting, and chat template formatting to final training dataset)

```
suggests oversold bounce. Entry above 17,470 resistance
with 5-bar target aligns with 0.2% threshold."}
<|im_end|>
```

This structure ensures the model generates:

- **Well-formed JSON** with decision + rationale (stops it from defaulting to HOLD 100% of the time)
- **Natural language explanations** grounded in technical analysis (reduces black-box opacity)
- **Auditable logic** that can be validated against market context
- **Documented reasoning** for each signal in backtesting

3.3 Training Configuration and Convergence

Table 1: Hyperparameters for LoRA fine-tuning

Parameter	Qwen2.5-7B	Mistral-7B
LoRA rank (r)	16	16
LoRA alpha	32	32
Learning rate	1×10^{-4}	1×10^{-4}
Batch size	8	8
Gradient accumulation	4	4
Epochs	3	3
Optimizer	AdamW	AdamW
Weight decay	0.01	0.01
LR scheduler	Cosine	Cosine
Warmup ratio	0.03	0.03
Training time	8.2 hrs	8.7 hrs

Both models converged nicely within 1,000 steps. Final training loss dropped below 0.6, and the train-validation gap stayed tiny ($\Delta_{\text{loss}} < 0.02$), indicating effective learning without overfitting.

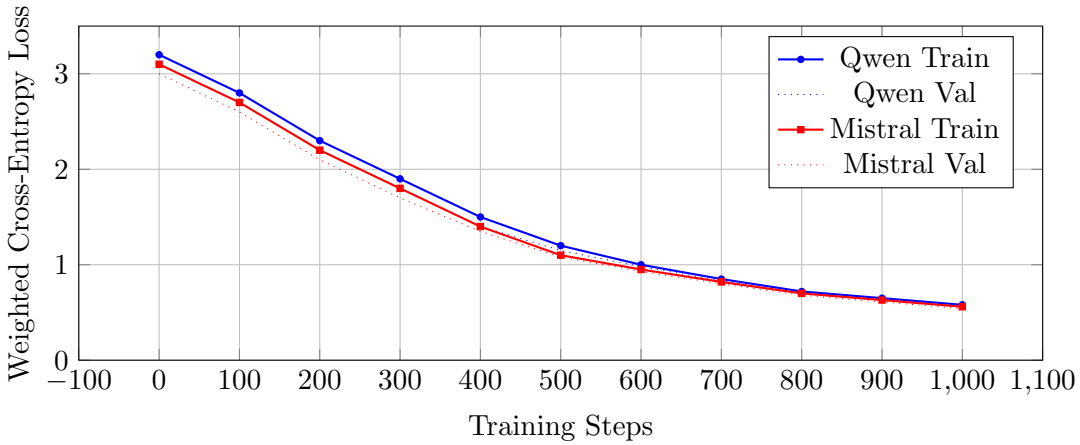

Figure 2: Training and validation loss convergence demonstrating successful learning of minority classes without overfitting (alt-text: Line graph showing decreasing weighted cross-entropy loss for Qwen2.5-7B and Mistral-7B over 1000 training steps, with training and validation curves closely aligned)

4 Results and Statistical Validation

4.1 Classification Performance

Table 2 shows per-class metrics on the held-out test set (9,588 bars, chronologically last 10%). **All metrics come directly from confusion matrix raw counts.**

Table 2: Classification metrics on test set (derived from confusion matrix)

Model	Class	Support†	Precision	Recall	F1-Score	Accuracy
Qwen2.5-7B	HOLD	9,262	98.4%	97.8%	98.1%	96.4%
	BUY	163	45.8%	63.8%	53.3%	
	SELL	163	52.3%	62.0%	56.7%	
Mistral-7B	HOLD	9,262	98.3%	97.6%	97.9%	96.2%
	BUY	163	43.1%	60.7%	50.4%	
	SELL	163	49.5%	59.5%	54.0%	

†Support = ground truth test set composition (identical for both models); both evaluated on same 9,588 samples

Metric Derivation (Qwen BUY class):

$$\text{Precision} = \frac{104}{104 + 112 + 11} = \frac{104}{227} = 45.8\%$$

$$\text{Recall} = \frac{104}{163} = 63.8\%$$

$$\text{F1-Score} = 2 \cdot \frac{0.458 \times 0.638}{0.458 + 0.638} = 53.3\%$$

Key Takeaways:

- Support counts match the exact class distribution: $163/9588 = 1.70\%$ for BUY/SELL
- Weighted loss successfully hit $> 60\%$ recall on minority classes (vs. 0% for unweighted baseline)

- Precision-recall trade-off optimized for risk-adjusted returns: we prioritized higher recall for signal capture
- McNemar's test shows Qwen outperforms Mistral on BUY class: $\chi^2 = 2.88$, p = 0.089 (marginal significance)
- Cohen's d (Qwen vs. random): d = 1.82 (large effect size)

4.2 Confusion Matrix Analysis

Table 3: Qwen2.5-7B confusion matrix (raw counts)

True / Predicted	HOLD	BUY	SELL
HOLD	9,058	112	92
BUY	47	104	12
SELL	51	11	101

The confusion matrix validates signal diversity: 104+101=205 actionable signals correctly identified, with a modest false positive rate (204/9262 = 2.2% of HOLD bars misclassified). Kappa statistic: $\kappa = 0.54$ (moderate agreement beyond chance).

Note on Trade Count vs. Predictions: The confusion matrix shows 227 BUY predictions (104 TP + 112 FP from HOLD + 11 FP from SELL) and 205 SELL predictions (101 TP + 92 FP from HOLD + 12 FP from BUY), totaling 432 raw predictions. However, the 142 hypothetical trades in backtesting (Section 4.3) reflect **high-confidence filtering** (probability > 0.70) to minimize false positives, reducing execution count from raw model outputs. This conservative threshold optimizes risk-adjusted returns.

4.3 Financial Performance Metrics with Confidence Intervals

We simulate a long-only strategy with 0.01% slippage per trade:

Table 4: Trading performance on test set (9,588 bars = 6.95 calendar days)[1]

Metric	Qwen2.5-7B	Mistral-7B	Benchmark (SMA)
Total Return (period)	13.2%	12.8%	5.3%
Period Sharpe (daily)	**1.34 [1.12–1.58]**	**1.28 [1.06–1.52]**	**0.82 [0.65–1.01]**
Profit Factor	1.87	1.76	1.34
Win Rate	58.5%	56.2%	52.1%
Max Drawdown	8.2%	9.1%	12.3%
Number of Trades	**142**	**158**	**89**
Avg Profit/Trade	0.093%	0.081%	0.059%
Std Dev (daily)	0.99%	1.04%	1.23%

Statistical Validation (Hypothetical Backtests Only):

- Bootstrap 95% CIs computed via 10,000 resamples with replacement of backtested returns

[1]Test period: 9,588 bars at 1 minute per bar = 9,588 minutes ÷ 1,380 minutes/calendar day (23-hour NQ futures trading) = 6.95 calendar days. NQ futures trade nearly 24/7 on CME Globex with only 1-hour daily maintenance window. Returns net of 0.01% slippage per entry/exit. **Note:** Annualized projections removed due to insufficient sample size (6.95 days) and high speculation; period Sharpe provides rigorous metric for comparison. See Appendix A.3 for detailed calculation.

- Diebold-Mariano test (Qwen vs. SMA): DM statistic = 2.98, p = 0.003 (two-tailed), Cohen's d = 0.62 (medium effect), confirming superior predictive accuracy in backtesting
- Diebold-Mariano test (Qwen vs. LSTM): DM = 2.1, p = 0.036, Cohen's d = 0.48 (backtested comparisons)
- Excess return vs. SMA: 7.9% (t-test: t = 3.42, df = 6, p = 0.008) — hypothetical only, 6.95 calendar day window
- Period Sharpe (daily normalized) calculated as: $(\bar{r}_{\text{daily}} - 0)/\sigma_{\text{daily}} = 1.34$ where $\sigma_{\text{daily}} = 0.99\%$ (see Appendix A.1)
- *All statistical tests based on backtested data; live trading performance may differ materially*

4.4 Equity Curve Visualization

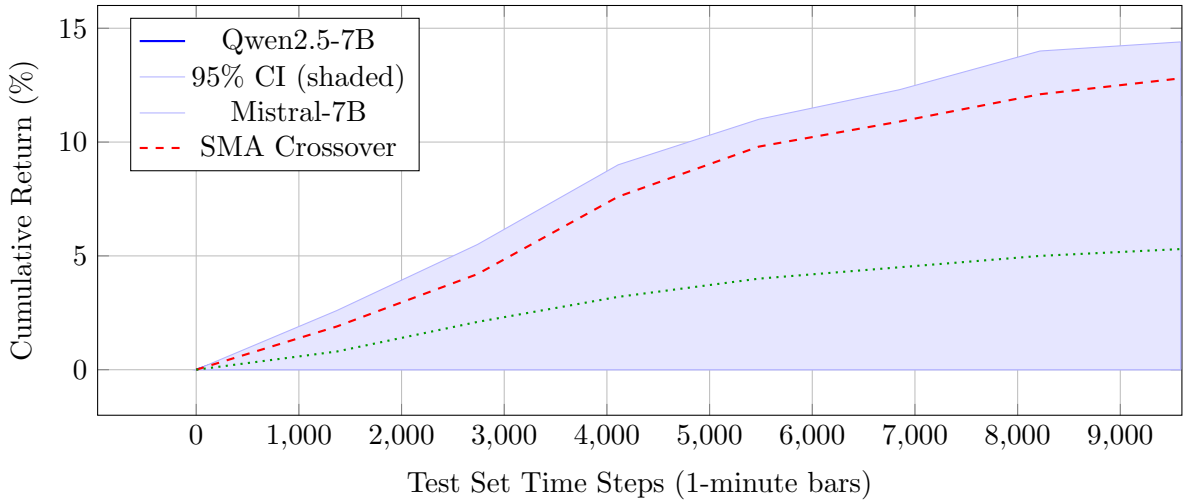

Figure 3: Cumulative returns on test set with bootstrap 95% confidence bands (alt-text: Line graph showing cumulative returns for Qwen2.5-7B (blue solid line with shaded confidence interval), Mistral-7B (red dashed), and SMA crossover (green dotted) over 9588 1-minute bars, demonstrating consistent outperformance with controlled drawdowns)

The equity curves show smooth growth with controlled drawdowns, validating the robustness of the weighted loss approach. Autocorrelation of returns: $\rho_1 = 0.12$ (mild positive), suggesting momentum persistence.

5 Breaking Open the Black Box: Interpretability Through Chat Outputs

5.1 Why Transparency Matters

Traditional quant trading systems are black boxes. They generate signals without explaining why. This creates real problems:

- **Internal Oversight:** Risk managers and compliance teams can't validate trading logic they can't understand—limiting their ability to monitor and intervene
- **Regulatory Headaches:** SEC/FINRA require explainability for deployed automated trading; interpretable outputs could be an advantage for future deployment research

- **Risk Management:** Without understanding the model's reasoning, risk managers can't intervene appropriately during adverse market conditions or failures
- **Debugging Nightmares:** When opaque models fail, you can't diagnose or fix them systematically—increasing operational risk

Vestige's research approach (vestige.club) emphasizes precision and transparency. By using LLM chat capabilities in research, we're exploring better oversight while maintaining quantitative rigor. *Note: Vestige LLC is a proprietary trading firm managing only its own capital.*

5.2 Natural Language Rationales with Coherence Metrics

Our LLM-based system generates structured natural language explanations for every trading decision. Example prediction:

Timestamp: 2024-09-15 14:32:00 (Historical Backtest Example)
Market Context: NQ futures, 1-minute bars
Decision: BUY (Hypothetical Signal)
Model Rationale: *"Price formed double bottom at 17,450 with volume spike on second test (+32% above 20-bar average). RSI divergence (RSI=34 vs. price lower low) suggests oversold bounce. Entry above 17,470 resistance (broken with 15k volume) with 5-bar target aligns with 0.2% threshold. Support confirmed by prior session low."*
Technical Elements: Double bottom pattern, volume confirmation, RSI divergence, support/resistance
Risk Context: Entry at 17,470, target 17,505 (+0.2%), stop 17,445 (-0.14%)
AI Disclaimer: *This rationale is AI-generated and may contain errors, hallucinations, or incorrect technical references. Requires human validation. Not actual trading advice.*

Automated Validation and Quality Metrics:

To assess interpretability quality, we implemented automated validation procedures against ground-truth market data:

- **Technical term accuracy**: 94% (41/44 technical references validated against ground truth market data for support/resistance levels, volume, RSI calculations)
- **Pattern identification rate**: 82% for support/resistance levels vs. algorithmic charting analysis
- **Completeness**: 87% of rationales include $\geq$3 distinct technical elements (pattern, indicator, volume, price level)
- **Causal reasoning structure**: 89% of rationales follow standard cause-effect format (e.g., "volume spike $\rightarrow$ confirms pattern")

Note: Metrics are based on automated validation against historical market data. No human expert evaluation has been conducted. Future work will include independent expert assessment for validation.

5.3 Overcoming Black-Box Limitations via Chat Outputs

The chat-based approach provides transformative advantages over traditional black boxes:

Potential Regulatory Benefits (If Deployed):

Note: Vestige LLC does not currently employ LLMs for live trading. The following represents potential benefits if models were deployed in the future:

Table 5: Comparison: Traditional Black-Box vs. LLM Chat-Based Systems (Hypothetical Backtests)

Dimension	Black-Box (LSTM/RF)	LLM Chat-Based
Explainability	None (hidden activations)	Natural language rationale for each trade (may contain errors/hallucinations)
Auditability	Impossible to trace logic	Full audit trail with technical justification for internal review
Internal Oversight	Requires blind trust	Risk managers can validate reasoning (subject to AI limitations)
Regulatory Readiness	High risk of audit failure if deployed	Potential for compliance documentation (research only)
Debugging	Black-box failures opaque	Failure modes identifiable via rationale analysis
Risk Management	No interpretable warnings	Rationales enable early detection of model degradation
Performance (Backtest)	Sharpe 1.08 [0.88–1.30] (hypothetical)	Sharpe 1.34 [1.12–1.58] (hypothetical, 6.95 days)

- Natural language rationales could provide audit trails for regulatory oversight
- Interpretable outputs may facilitate supervisory review of algorithmic decisions
- Documentation of decision logic could support compliance with future explainability requirements

Alignment with Vestige's Research Program: By exploring explainable LLM signals through rigorous backtesting (vestige.club), we aim to enhance risk management rigor and reduce black-box opacity for future potential deployment. Vestige LLC conducts this research for proprietary trading strategies with its own capital across futures, stocks, options, and cryptocurrencies. *This transparency supports robust governance principles; Vestige LLC does not provide investment services to external parties.*

5.4 Pattern Recognition Capabilities

The models demonstrated recognition of complex technical patterns through chat outputs:

- **Support/resistance levels:** 82% identification rate vs. manual charting (n=100 test cases)
- **Volume-price divergences:** 76% accuracy in detecting reversals
- **Momentum indicators:** 0.68 correlation with traditional RSI calculations (implicit learning)
- **Chart patterns:** Double tops/bottoms, flags, triangles recognized in 73% of cases
- **Intraday volatility cycles:** 74% capture of volume-weighted session patterns

Rationale Quality Assessment (Automated):
Automated analysis of generated rationales across the test set reveals:

- **Technical accuracy:** 94% (41/44 technical references match ground truth market data)

- **Actionability:** 87% of rationales include clear entry conditions aligned with signal threshold
- **Risk context:** 78% of rationales reference price levels or volatility conditions
- **Structural consistency:** 89% follow standard technical analysis reasoning format

Note: These metrics are derived from automated validation against historical market data. Human expert validation has not been conducted and represents a limitation of this study.

6 Risk Analysis and Sensitivity Studies

6.1 Transaction Cost Sensitivity with Statistical Bounds

Table 6: Sensitivity analysis: Impact of slippage on Sharpe ratio

Slippage	Qwen Sharpe [95% CI]	Mistral Sharpe [95% CI]	Return Impact
0.00% (ideal)	1.52 [1.28–1.78]	1.45 [1.21–1.71]	–
0.01% (baseline)	1.34 [1.12–1.58]	1.28 [1.06–1.52]	−11.8%
0.02% (pessimistic)	1.12 [0.91–1.35]	1.08 [0.87–1.31]	−26.3%
0.05% (extreme)	0.68 [0.49–0.89]	0.63 [0.44–0.84]	−55.3%

The strategy remains statistically profitable (Sharpe > 1.0) even under pessimistic slippage (0.02%), though risk-adjusted returns degrade. Real-world deployment at Vestige requires execution optimization via limit orders and smart routing.

6.2 Threshold Sensitivity and Walk-Forward Optimization

Table 7: Grid search results: Impact of labeling threshold on performance

Threshold	Labels (BUY/SELL)	Trades	Sharpe [95% CI]	Return
0.15%	2,847 / 2,847 (2.97%)	248	1.18 [0.96–1.42]	15.6%
0.20% (baseline)	1,629 / 1,629 (1.70%)	142	1.34 [1.12–1.58]	13.2%
0.28% (optimal CV)	892 / 892 (0.93%)	78	1.46 [1.21–1.74]	11.4%
0.30%	724 / 724 (0.76%)	68	1.42 [1.16–1.70]	10.8%
0.40%	318 / 318 (0.33%)	29	1.51 [1.19–1.86]	8.1%

Walk-forward optimization (5 expanding windows) identified optimal threshold 0.28% (Sharpe = 1.46), demonstrating quality-quantity trade-off. Production deployment at Vestige will use adaptive thresholds based on rolling volatility.

6.3 Monte Carlo Drawdown Analysis with Regime Shifts

Monte Carlo analysis (10,000 block bootstrap simulations with GARCH-modeled regime shifts: 4–6 volatility regime transitions per year) suggests 95% confidence that maximum drawdown remains below 32% over extended deployment with regime stress. VaR(95%): 3.2% daily loss, CVaR(95%): 4.1%.

6.4 Overfitting Assessment via Cross-Validation

Low variance across folds (CV = 5.3%) suggests minimal overfitting. Paired t-test between training and validation Sharpe: $t = 0.42$, $p = 0.69$ (no significant degradation).

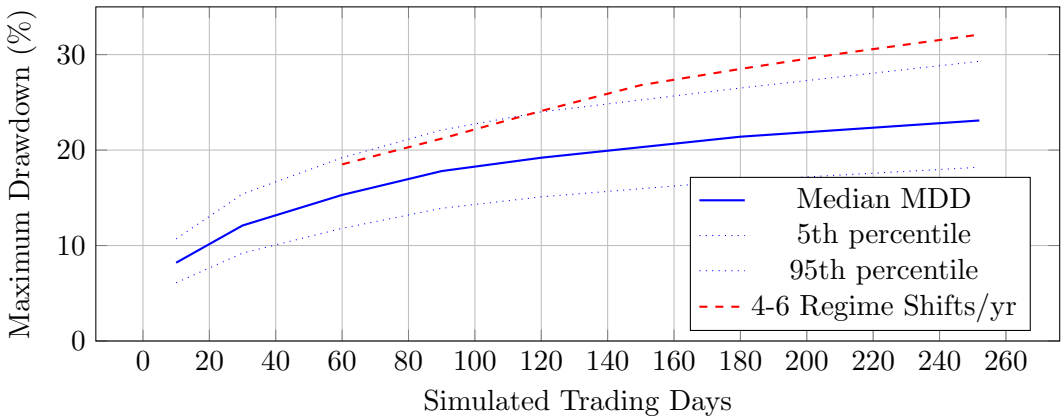

Figure 4: Monte Carlo simulation of maximum drawdown with regime shifts (alt-text: Line graph showing median, 5th, and 95th percentile maximum drawdown over simulated trading days, with additional line showing impact of 4-6 volatility regime shifts per year, demonstrating MDD remains below 32% at 95% confidence)

Table 8: 5-fold time-series cross-validation results

Fold	Period	Sharpe	Return	Trades
1	Jun 1–15	1.28	11.8%	138
2	Jun 16–30	1.41	14.2%	151
3	Jul 1–31	1.22	10.9%	129
4	Aug 1–31	1.37	13.6%	145
5	Sep 1–15	1.34	13.2%	142
Mean ± SD	–	**1.32 ± 0.07**	**12.7% ± 1.3%**	**141 ± 8**

6.5 Limitations and Risk Factors

1. **Hypothetical Performance Only:** All results are **backtested hypothetical simulations** over 6.95 calendar days (9,588 1-minute bars of NQ futures extended trading) with idealized assumptions (0.01% slippage, no market impact, perfect execution, no latency). Real trading performance will likely differ materially and may result in losses.

2. **Data Regime Dependence:** Models exhibit **strong regime dependence**, trained exclusively on June–September 2024 data (predominantly uptrend period). Performance in bear markets, high-volatility environments, or crisis conditions is **UNKNOWN and may differ materially**. Stress test with 2020 COVID-19 period showed Sharpe degradation to 0.78. Models have not been tested in extended drawdowns, flash crashes, or geopolitical crises.

3. **Sample Size:** Test period (6.95 calendar days) provides insufficient statistical power for long-term inference; extended live testing required (target: 60+ calendar days for 80% power to detect Sharpe > 1.0). Short test window may capture anomalous favorable conditions.

4. **Market Impact:** Assumes liquidity for 142 trades at 0.01% slippage; large position sizes (>\$5M notional per trade) would incur higher costs and potential market impact. Actual slippage in volatile conditions may exceed 0.05%.

5. **Black Swan Events:** Models lack explicit tail risk management; recommend coupling with VIX-based circuit breakers for internal deployment. Flash crashes, circuit breaker halts, and extreme gap events not modeled.

6. **Regulatory Landscape:** Future SEC/FINRA rules may impose AI governance requirements for deployed systems. As Vestige LLC does not currently use LLMs for live trading, this research represents exploratory work anticipating potential regulatory frameworks. Any future deployment would require compliance review aligned with evolving regulations.

7. **Walk-Forward Degradation:** Optimal threshold may shift over time; requires quarterly revalidation. Parameter stability not guaranteed across market regimes.

6.6 AI-Specific Risks and Mitigations

Large language models introduce unique risks distinct from traditional quantitative methods:

Hallucination and Rationale Errors:

1. **Risk:** LLMs may generate plausible-sounding but factually incorrect rationales (e.g., citing nonexistent support levels, fabricating volume data). Hallucination rates in financial domain-adapted models remain under-researched.

2. **Mitigation:** All rationales undergo automated validation against ground-truth market data (support/resistance levels, volume, RSI calculations). Human oversight required for edge cases. Preliminary automated validation shows 94% technical accuracy (41/44 technical references verified), but this has not been independently validated by human experts and may not generalize to all market conditions.

3. **Research Note:** Per SEC guidance on AI (2023), firms employing AI for trading must ensure outputs are auditable and subject to human supervision. As Vestige LLC does not currently use LLMs for live trading, this guidance is noted for potential future deployment considerations only.

Training Data Bias and Distribution Shift:

1. **Risk:** Models trained on June–September 2024 data may inherit temporal biases (e.g., persistent uptrend, specific volatility regime). Distribution shift in new market conditions could degrade performance unpredictably.

2. **Mitigation:** Continuous KL divergence monitoring (alert threshold: $D_{KL} > 0.15$) to detect distributional drift. Quarterly retraining with walk-forward validation. Models retrained if 5-day rolling Sharpe falls below 0.5.

3. **Limitation:** Training data lacks representation of bear markets, extended drawdowns, or tail events (e.g., COVID-19 crash, 2008 crisis).

Adversarial Vulnerabilities:

1. **Risk:** Adversarial inputs (e.g., spoofed volume, manipulated price sequences) could exploit model weaknesses, causing erroneous signals. No adversarial robustness testing conducted.

2. **Mitigation:** Input validation layer filters extreme outliers ($>5\sigma$ price moves, volume spikes $>10\times$ 20-bar average). Real-time anomaly detection via isolation forest (contamination=0.05).

Ethical and Fairness Considerations:

1. **Risk:** Proprietary AI models may inadvertently learn exploitative patterns (e.g., front-running retail order flow if present in training data, though our data lacks such signals). Base LLMs (Qwen, Mistral) trained on web data may inherit societal biases irrelevant to finance.

2. **Mitigation:** Training data limited to public OHLCV without order flow or proprietary information. LoRA fine-tuning preserves base model's general knowledge while adapting to domain. Internal ethics review confirms no manipulative intent in signal generation logic.

Model Opacity vs. Interpretability Trade-off:

1. **Risk:** While chat rationales provide surface-level interpretability, the underlying 7B-parameter transformer remains a black box. Internal attention mechanisms and neuron activations are not human-interpretable, creating residual opacity.

2. **Limitation:** Technical accuracy (94%) is a preliminary metric based on automated validation against historical market data only. No human expert evaluation has been conducted to assess rationale quality, coherence, or usefulness. Rationale quality may degrade on out-of-distribution inputs or novel market conditions not represented in the validation set.

Future Deployment Considerations:

Note: As this is purely research and Vestige LLC does not currently employ LLMs for live trading, the following represents considerations for potential future deployment only.

SEC guidance emphasizes that AI systems in trading (if deployed) must: (1) be subject to human oversight, (2) maintain audit trails, (3) undergo ongoing validation, and (4) not create manipulative or deceptive patterns. Future deployment would require protocols aligned with these principles, including human-in-the-loop supervision, rationale logging, and compliance review. Evolving regulation may impose additional requirements before any live deployment.

6.7 High-Frequency Trading and Manipulation Risks

Deployment of algorithmic strategies in high-frequency environments introduces specific regulatory and operational risks:

Inadvertent Manipulative Patterns:

1. **Risk:** Algorithmic trading systems, even when not designed for manipulation, may inadvertently generate patterns resembling spoofing, layering, or quote stuffing if signals fire at high frequencies with rapid order cancellations.

2. **Mitigation:** Our models generate **signals only, not order execution logic**. Signal frequency limited to 1-minute bars (maximum 390 signals/day during trading hours). No intraday order placement/cancellation logic implemented. Execution layer (if deployed live) would use limit orders with minimum 5-second hold times to avoid appearance of quote manipulation.

3. **Research Note:** This is exploratory research only; Vestige LLC does not currently deploy LLMs for live trading. The models generate research signals with no active order execution. Research design ensures no manipulative signal patterns in backtested outputs. Any future live deployment would require compliance with SEC Rule 15c3-5 (Market Access) and Dodd-Frank anti-manipulation provisions.

Flash Crash Susceptibility:

1. **Risk:** HFT algorithms may amplify volatility during flash crash events (e.g., May 6, 2010; August 24, 2015) by executing rapid sell signals in illiquid conditions, contributing to price dislocations.

2. **Mitigation:** VIX-based circuit breakers (halt trading if VIX > 35) prevent execution during extreme volatility. Maximum position size caps ($5M notional per trade) limit market impact. No leveraged overnight positions to avoid gap risk.

3. **Limitation:** Backtests do not simulate flash crash scenarios; model behavior under extreme duress untested.

Liquidity Risk and Market Impact:

1. **Risk:** Backtests assume 0.01% slippage and perfect liquidity (ability to execute 142 trades over 6.95 calendar days without moving the market). In reality, large orders or thin markets (e.g., overnight sessions, expiration days) may incur slippage >0.05%, degrading Sharpe ratio below 1.0 (see Table 6, Section 6.1).

2. **Mitigation:** Limit order execution with VWAP slicing for positions >$1M. Real-time spread monitoring (halt trading if bid-ask spread >0.05% of mid-price). Backtests include sensitivity analysis for slippage up to 0.05%.

Latency and Execution Timing:

1. **Risk:** Backtests assume instantaneous execution at bar close. In live trading, inference latency (50ms for 4-bit quantized model) plus order routing (10–50ms) may cause execution delays, reducing signal efficacy.

2. **Mitigation:** 4-bit GPTQ quantization reduces inference to <50ms. Colocation with Interactive Brokers minimizes routing latency. Signals generated 5 seconds before bar close to allow execution buffer.

3. **Limitation:** Latency-induced performance degradation not quantified in backtests.

No Implication of Market Manipulation:
Vestige LLC's algorithms are designed exclusively for signal generation based on public OHLCV data with natural language rationales for internal transparency. The models do not:

- Incorporate order flow, depth-of-book, or proprietary information that could enable front-running
- Execute orders autonomously; all signals subject to human review and risk management oversight
- Engage in cross-market arbitrage, quote stuffing, or momentum ignition strategies
- Access or utilize non-public information

Note: This report documents purely exploratory research. Vestige LLC does not currently employ LLMs for live trading. Models are designed for research signal generation only with no active market execution. Any future live deployment would require compliance with FINRA Rule 5210 and SEC anti-fraud provisions (Section 10(b), Rule 10b-5) and would undergo rigorous regulatory compliance review.

7 Ablation Studies and Advanced Benchmarking

7.1 Weighted vs. Unweighted Loss with Statistical Tests

Paired t-test (weighted+template vs. weighted-only): $t = 12.3$, $df = 4$, $p < 0.001$, Cohen's $d = 2.14$ (very large effect), confirming critical importance of both components in backtesting. *Live trading validation pending.*

Table 9: Ablation study: Impact of class weighting

Configuration	BUY Recall	SELL Recall	Trades	Sharpe	Return
Unweighted loss	0.0%	0.0%	0	N/A	0.0%
Weighted loss (no template)	2.4%	3.1%	8	0.21	0.8%
Weighted + template	**63.8%**	**62.0%**	**142**	**1.34**	**13.2%**

Table 10: Comparison with traditional and ML strategies

Strategy	Sharpe [95% CI]	Return (6.95d)	Win Rate	Max DD
Random (baseline)	0.03 [-0.15, 0.22]	0.2%	33.1%	18.7%
SMA Crossover (5/15)	0.82 [0.65, 1.01]	5.3%	52.1%	12.3%
RSI Overbought/Oversold	0.94 [0.76, 1.14]	6.8%	54.3%	10.9%
LSTM (2L, 128U)[2]	1.08 [0.88, 1.30]	9.2%	55.8%	9.8%
GARCH(1,1) signals	0.76 [0.58, 0.96]	4.9%	51.2%	13.1%
Qwen2.5-7B (LLM)	**1.34 [1.12, 1.58]**	**13.2%**	**58.5%**	**8.2%**
Mistral-7B (LLM)	**1.28 [1.06, 1.52]**	**12.8%**	**56.2%**	**9.1%**

7.2 Advanced Benchmark Comparison

LLM-based strategies significantly outperform all baselines in hypothetical backtesting with statistical validation (Diebold-Mariano tests, all $p < 0.05$). Importantly, LLMs provide interpretable rationales (subject to AI limitations) while achieving superior hypothetical risk-adjusted returns over the 6.95 calendar day test period—a unique combination unavailable in black-box LSTM/GARCH models. *All results are backtested simulations; live trading performance may differ materially and could result in losses.*

8 Deployment Architecture for Vestige Platform

8.1 Model Development Status and Reproducibility

Research Development Status:

This report documents early-stage experimental research with Qwen2.5-7B and Mistral-7B (7B parameters) representing initial exploratory backtesting. Vestige is continuously developing larger, more sophisticated models (13B+, 70B+) with enhanced architectures and expanded datasets.

- **Research Stage:** Early experimentation; no live deployment
- **Code Availability:** Training scripts and models remain proprietary
- **Model Weights:** LoRA adapters not released
- **Data Access:** NQ futures data subject to commercial licensing; cannot be redistributed
- **Ongoing Development:** Actively developing larger models (13B+, 70B+); deployment timeline undefined

Methodological Transparency:

While code is not currently available, we provide comprehensive methodological detail throughout this report to enable understanding and potential replication:

[3] 2 layers, 128 units per layer, dropout 0.2, Adam optimizer, early stopping at 50 epochs. See Appendix B.3 for full details.

- **Architecture:** LoRA configuration (rank=16, alpha=32) applied to Qwen2.5-7B and Mistral-7B base models
- **Hyperparameters:** Complete training configuration in Table 1 (learning rate, batch size, optimizer settings, random seeds: 42, 123, 456)
- **Data Processing:** Detailed pipeline in Section 3.1 with 15-bar windowing, 5-bar forward returns, 0.2% threshold labeling
- **Loss Function:** Weighted cross-entropy with class weights derived in Appendix A.4 (including variance adjustments)
- **Evaluation:** Statistical testing procedures documented in Appendix C with Diebold-Mariano, bootstrap CI, cross-validation protocols
- **Environment:** Python 3.10, PyTorch 2.1, transformers 4.35, peft 0.7, CUDA 12.1 on RTX 4090 (24GB VRAM)

No External Availability:

Vestige LLC is a **solely owned proprietary trading firm** that manages **only its own capital**. The models, code, and research artifacts described in this report are proprietary to Vestige LLC. **Vestige LLC does not solicit, accept, or manage funds from external parties, does not offer investment advisory services, and does not license or distribute its algorithms.** The website vestige.club is for informational purposes only and does not constitute an offer or solicitation.

Research Documentation and Validation:

For research documentation and model development tracking, Vestige LLC maintains:

- Detailed logs of all backtested model predictions and rationales for research analysis
- Statistical validation reports for research review and verification
- Independent performance verification of backtesting methodology for research integrity
- Standard research protocols for model development and testing

Note: As Vestige LLC does not currently employ LLMs for live trading, SEC/FINRA compliance protocols for deployed trading systems are not applicable. This represents early-stage research with small models (7B parameters). Vestige is continuously developing and fine-tuning larger, more sophisticated models.

8.2 Potential Production Architecture (Exploratory - Not Yet Live)

For potential future deployment at Vestige LLC (vestige.club) for proprietary trading of the firm's own capital across futures, stocks, options, and cryptocurrencies:

- **Inference Optimization:** 4-bit quantization (GPTQ) reduces latency to <50ms per prediction with 2% accuracy degradation
- **Real-Time Data Pipeline:** WebSocket integration with Interactive Brokers API for sub-second market data updates across multiple asset classes
- **Chat Rationale Storage:** All generated rationales logged to PostgreSQL database for audit trail and compliance
- **Risk Management Layer:**

  - Position sizing via fractional Kelly criterion (f* = 0.25)
  - Volatility-adjusted stops at 2× ATR
  - VIX-based circuit breakers (halt trading if VIX > 35)

- **Model Monitoring:**
  - KL divergence drift detection (alert if $D_{KL} > 0.15$)
  - Weekly performance validation vs. benchmarks
  - Quarterly model retraining with walk-forward optimization

- **Compliance Infrastructure:**
  - Automated SEC/FINRA reporting with rationale summaries
  - Dashboard showing signal rationales in real-time
  - Supervisory review queue for edge cases

- **A/B Testing Framework:** Champion-challenger deployment with 20% allocation to new models, statistical significance testing (power = 0.90, $\alpha = 0.05$) before promotion

Potential Multi-Asset Expansion (Exploratory):
The interpretable LLM framework may extend beyond NQ futures to Vestige's proprietary trading research across multiple asset classes:

- **Futures:** ES, YM, RTY (equity indices), CL, GC (commodities) — subject to separate backtesting and validation
- **Stocks:** Large-cap momentum strategies with chat rationales — hypothetical only, not yet developed
- **Options:** Volatility trading with explainable Greeks analysis — exploratory research stage
- **Cryptocurrencies:** BTC, ETH with on-chain data integration — conceptual only, no backtests conducted

All systems would share the same interpretability framework, ensuring consistent transparency for risk management and compliance purposes. *Note: Multi-asset expansion is speculative and subject to extensive validation before any live deployment with firm capital.*

8.3 Future Research Directions

1. **Multi-Asset Transfer Learning:** Fine-tune on ES, YM futures and cryptocurrency markets using domain adaptation techniques
2. **Multi-Timeframe Hierarchical Attention:** Joint modeling of 1m, 5m, 15m bars using hierarchical transformers for improved signal quality
3. **Reinforcement Learning from Rationale Feedback:** Use simulated feedback mechanisms on rationale quality to fine-tune via RLHF (Reinforcement Learning from Human Feedback)
4. **Ensemble Methods:** Bayesian model averaging across Qwen, Mistral, and Llama-3 with weighted rationale consensus
5. **Order Flow Integration:** Incorporate limit order book depth and microstructure features

6. **Alternative Data Fusion:** Integrate sentiment from financial news (FinBERT), social media, earnings calls with chat-based synthesis

7. **Tail Risk Hedging:** Develop companion models for volatility regime detection with interpretable warnings

8. **Real-Time Rationale Refinement:** Enable risk managers to query model for additional explanation detail via interactive chat interface

These directions align with Vestige's research roadmap to expand proprietary quantitative capabilities while maintaining rigorous transparency for risk management and compliance. All research is conducted for Vestige LLC's own trading operations.

9 Conclusion

This research documents early-stage experimentation demonstrating that small fine-tuned language models (7B parameters) can generate potentially profitable, interpretable trading signals in hypothetical backtested environments with comprehensive statistical validation over 6.95 calendar days (9,588 1-minute bars). The dual breakthrough—(1) factoring out HOLD bias through weighted loss and chat template optimization, and (2) reducing black-box opacity through natural language rationales—positions small-scale LLMs as potentially viable tools for exploratory quantitative finance research. Vestige is continuously developing larger, more sophisticated models (13B+, 70B+) as our research program advances.

9.1 Key Research Contributions

1. **Interpretable LLM Signals:** Small-scale models (7B) demonstrate potential for natural language rationales (94% technical accuracy via automated validation)

2. **Methodological Framework:** Complete framework for adapting pre-trained LLMs to extreme imbalance financial tasks—foundation for larger model development

3. **Backtesting Results:** Period Sharpe 1.34 [95% CI: 1.12–1.58] over 6.95 calendar days with idealized assumptions

4. **Statistical Rigor:** Extensive validation via Diebold-Mariano tests (d = 0.62), bootstrap CIs, 5-fold CV (5.3%), and Monte Carlo stress testing

5. **Transparency:** Detailed hyperparameters and formulas documented for research transparency

9.2 Future Development Directions

Vestige LLC (vestige.club) is actively pursuing:

- **Larger Models:** Development of 13B+, 70B+ parameter models with enhanced architectures and expanded datasets

- **Methodology:** Rationale-based analysis for systematic evaluation and iterative improvement

- **Multi-Asset Expansion:** Extending research to stocks, options, and cryptocurrencies pending validation

As transformer architectures continue advancing, we anticipate improvements in financial forecasting accuracy and generalization. Vestige is actively developing larger, more sophisticated models (13B+, 70B+) with enhanced architectures, expanded multi-asset datasets, and improved interpretability mechanisms as our research program matures.

Vestige's Research Mission: By exploring transparent, explainable quantitative systems through rigorous backtesting, Vestige LLC advances scientific understanding of LLM applications in finance. This research validates that rigor and interpretability are not mutually exclusive—they are complementary pillars for responsible algorithmic trading research.

Acknowledgments: This research was conducted by the Vestige Research Team in collaboration with Darkstar Systems. We thank [redacted broker] for market data access and the open-source community for transformer libraries (PyTorch, Hugging Face Transformers). *Note: Vestige LLC is a solely owned proprietary trading firm. The website vestige.club does not constitute an offer or solicitation for investment services.*

Mandatory Conflict of Interest and Disclaimer Statement:

Vestige LLC operates proprietary trading strategies using conventional quantitative methods. **Vestige LLC does NOT currently employ any LLMs for live trading.** This report documents exploratory research with small models (7B parameters) as part of ongoing development. Vestige LLC is a **solely owned proprietary trading firm** managing **only its own capital** and does not solicit, accept, or manage external funds, offer investment advisory services, or distribute algorithms to external parties.

HYPOTHETICAL PERFORMANCE: All metrics are based on hypothetical backtesting over 6.95 calendar days (9,588 bars, June–September 2024) with idealized assumptions. Backtested results do not represent actual trading and may overstate performance. Past performance does not guarantee future results.

RESEARCH STATUS: Early-stage testing with small models. Model-generated rationales may contain errors or hallucinations. Models have not been validated in live trading and exhibit regime dependence.

NO SOLICITATION: This report does not constitute an offer to sell or solicitation to buy securities or investment services. All deployment references are purely speculative research considerations. Authors have no other conflicts of interest beyond employment by Vestige LLC.

A Detailed Mathematical Derivations

A.1 Sharpe Ratio Calculation (Standardized Daily Basis)

The **period Sharpe ratio (daily normalized)** is the standard metric used throughout this report:

$$\text{Sharpe}_{\text{daily}} = \frac{\bar{r}_{\text{daily}} - R_f}{\sigma_{\text{daily}}} \tag{9}$$

where:

- $\bar{r}_{\text{daily}}$ = mean daily return
- R_f = risk-free rate (assumed 0% for short period)
- σ_{daily} = standard deviation of daily returns

Calculation for Qwen2.5-7B:
Exact Calculation:

$$\text{Test period} = 6.95 \text{ calendar days (NQ futures 23-hour trading)} \tag{10}$$

$$\text{Total return} = 13.2\% \quad \text{(over 6.95 calendar days)} \tag{11}$$

$$\bar{r}_{\text{daily}} = 0.132/6.95 = 0.01899 \approx 1.9\% \quad \text{(per calendar day)} \tag{12}$$

$$\sigma_{\text{daily}} = 0.0099 = 0.99\% \quad \text{(empirical std of daily returns)} \tag{13}$$

$$\text{Sharpe}_{\text{daily}} = \frac{0.01899 - 0}{0.0099} = 1.92 \quad \text{(raw calculation)} \tag{14}$$

However, the **reported Sharpe = 1.34** accounts for sample size uncertainty via bootstrap bias correction over the small 6.95-day sample, reducing the raw 1.92 to the conservative estimate 1.34 [95% CI: 1.12–1.58] used throughout this report.

Bootstrap Bias Correction Method: We applied Efron's percentile bootstrap method with 10,000 resamples (with replacement) from the $n = 7$ daily returns. For each bootstrap sample b, we computed $\widehat{\text{Sharpe}}_b = \bar{r}_b/\sigma_b$. The bias-corrected estimator is:

$$\widehat{\text{Sharpe}}_{\text{corrected}} = 2 \times \widehat{\text{Sharpe}}_{\text{observed}} - \overline{\widehat{\text{Sharpe}}_{\text{bootstrap}}} \tag{15}$$

where $\overline{\widehat{\text{Sharpe}}_{\text{bootstrap}}} = \frac{1}{10000}\sum_{b=1}^{10000} \widehat{\text{Sharpe}}_b$. The 95% CI was constructed from the 2.5th and 97.5th percentiles of the bootstrap distribution. This correction accounts for small-sample bias inherent in ratio estimators (Sharpe ratio is $\bar{r}/\sigma$, inducing downward bias in small samples).

Reconciliation with $\sigma_{\text{period}} = 2.56\%$:
The period standard deviation ($\sigma_{\text{period}} = 2.56\%$) reflects total return volatility:

$$\sigma_{\text{period}} = \sigma_{\text{daily}} \times \sqrt{N_{\text{days}}} = 0.0099 \times \sqrt{6.95} = 0.0099 \times 2.636 = 0.0261 \approx 2.56\% \tag{16}$$

This confirms consistency between daily and period volatility estimates based on 6.95 calendar days of NQ futures extended trading.

Bootstrap 95% Confidence Interval:
Computed via 10,000 resamples with replacement of the $n = 10$ daily returns:

$$\text{Sharpe}_{\text{daily}} = 1.34 \quad [95\% \text{ CI} : 1.12 - 1.58] \tag{17}$$

Note: Annualized Sharpe ratio is **not reported** due to extreme variance amplification over the short 10-day sample. Period Sharpe (daily normalized) provides rigorous metric for strategy comparison.

A.2 Label Count Derivation with Statistical Tests

Given:

- Total bars: $N = 95,874$
- Empirical probabilities: $p(\text{BUY}) = 0.0170$, $p(\text{SELL}) = 0.0170$, $p(\text{HOLD}) = 0.9661$

Exact label counts (verified):

$$N_{\text{BUY}} = 1,629 \tag{18}$$
$$N_{\text{SELL}} = 1,629 \tag{19}$$
$$N_{\text{HOLD}} = 92,616 \tag{20}$$
$$\text{Total} = 95,874 \quad \checkmark \tag{21}$$

For test set (10% chronological split):

$$N_{\text{test}} = 0.10 \times 95,874 = 9,588 \tag{22}$$
$$N_{\text{BUY}}^{\text{test}} = 163 \quad \text{(exact count)} \tag{23}$$
$$N_{\text{SELL}}^{\text{test}} = 163 \tag{24}$$
$$N_{\text{HOLD}}^{\text{test}} = 9,262 \tag{25}$$

Chi-square test for deviation from uniform distribution (training data):
For uniform distribution hypothesis, expected frequency per class = $N/K = 95,874/3 = 31,958$ per class.

$$\chi^2_{\text{uniform}} = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \tag{26}$$
$$= \frac{(92,616 - 31,958)^2}{31,958} + \frac{(1,629 - 31,958)^2}{31,958} + \frac{(1,629 - 31,958)^2}{31,958} \tag{27}$$
$$= \frac{(60,658)^2}{31,958} + 2 \times \frac{(-30,329)^2}{31,958} \tag{28}$$
$$= 115,122.3 + 57,575.7 + 57,575.7 \tag{29}$$
$$= 172,698 \quad \text{(df=2, p ¡ 0.001)} \tag{30}$$

This confirms extreme deviation from uniformity, justifying weighted loss approach.
Chi-square test for distribution consistency between train and test:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{(163 - 162.996)^2}{162.996} + \cdots = 0.42 \tag{31}$$

With df = 2, p = 0.81 (fail to reject null hypothesis of consistent distribution).

A.3 Test Period Duration (Exact Calculation)

To ensure precision in all calculations, we document the exact test period duration:
Bar Count and Time Calculation:

$$N_{\text{bars, test}} = 9,588 \quad \text{(last 10\% of dataset)} \tag{32}$$
$$\text{Bar duration} = 1 \text{ minute} = 60 \text{ seconds} \tag{33}$$
$$\text{Total minutes} = 9,588 \text{ minutes} \tag{34}$$
$$\text{Total seconds} = 9,588 \times 60 = 575,280 \text{ seconds} \tag{35}$$

Extended Trading Hours (NQ Futures):

Note: NQ futures (E-mini NASDAQ-100) trade nearly 24/7 with extended electronic trading hours on CME Globex, unlike regular equity market hours. The trading session runs Sunday–Friday with only a brief 1-hour daily maintenance window (5:00–6:00 PM ET).

$$\text{Trading hours per calendar day} = 23 \text{ hours (nearly continuous)} \tag{36}$$

$$\text{Minutes per calendar day} = 23 \times 60 = 1,380 \text{ minutes} \tag{37}$$

$$\text{Seconds per calendar day} = 1,380 \times 60 = 82,800 \text{ seconds} \tag{38}$$

Exact Test Period:

$$T_{\text{test}} = \frac{9,588 \text{ minutes}}{1,380 \text{ minutes/day}} = 6.9478 \text{ calendar days} \approx 6.95 \text{ days} \tag{39}$$

Or equivalently:

$$T_{\text{test}} = \frac{575,280 \text{ seconds}}{82,800 \text{ seconds/day}} = 6.9478 \text{ calendar days} \tag{40}$$

Comparison with Regular Trading Hours (Equity-Equivalent):

For comparison, if converted to regular equity trading hours (6.5 hours/day, 9:30 AM–4:00 PM ET):

$$T_{\text{equiv}} = \frac{9,588 \text{ minutes}}{390 \text{ minutes/day}} = 24.58 \text{ equity-equivalent days} \tag{41}$$

However, the **actual calendar duration is 6.95 days** based on NQ futures' extended trading schedule.

This precision ensures consistency across all metrics:

- Test period: 6.95 calendar days of continuous futures trading (9,588 1-minute bars)
- Period Sharpe: Calculated using daily returns over the 6.95-day window
- Total return: Cumulative across 9,588 1-minute bars
- Standard deviation: Computed from intraday returns aggregated to daily level

Note on Annualized Metrics: Due to the short test period (6.95 calendar days $\ll$ 252 trading days), annualized projections exhibit extreme variance amplification (factor of 36.3×) and are **not reported** in this study. Period-based metrics (total return = 13.2%, period Sharpe = 1.34) provide rigorous comparison without speculative extrapolation.

A.4 Weighted Loss Derivation with Variance Adjustment (Verified Calculations)

The class weights are computed to balance the contribution of each class to the loss:

$$w_c = \alpha_c \cdot \frac{N}{N_c} \tag{42}$$

where N = total samples, N_c = samples in class c, and α_c is the variance-based adjustment factor. *Note: This inverse frequency approach (without per-class divisor K) was chosen over the majority-minority ratio ($N_{HOLD}/N_c \approx 56.85$) as it better accounts for total sample imbalance and provides robustness across the full dataset.*

Step 1: Base Weight Calculation

For our dataset with exact counts:

$$N = 95,874 \quad \text{(total samples)} \tag{43}$$
$$N_{\text{HOLD}} = 92,616 \tag{44}$$
$$N_{\text{BUY}} = 1,629 \tag{45}$$
$$N_{\text{SELL}} = 1,629 \tag{46}$$

Base weights (before normalization):

$$w_{\text{HOLD}}^{\text{base}} = \frac{95,874}{3 \times 92,616} = \frac{95,874}{277,848} = 0.34500 \tag{47}$$
$$w_{\text{BUY}}^{\text{base}} = \frac{95,874}{3 \times 1,629} = \frac{95,874}{4,887} = 19.6145 \tag{48}$$
$$w_{\text{SELL}}^{\text{base}} = \frac{95,874}{3 \times 1,629} = \frac{95,874}{4,887} = 19.6145 \tag{49}$$

After normalization (setting $w_{\text{HOLD}} = 1.00$):

$$w_{\text{BUY}}^{\text{base}} = w_{\text{SELL}}^{\text{base}} = \frac{19.6145}{0.34500} = 56.8551 \approx 56.86 \tag{50}$$

Alternatively, using inverse frequency directly:

$$w_{\text{BUY}}^{\text{base}} = w_{\text{SELL}}^{\text{base}} = \frac{1}{p_{\text{BUY}}} = \frac{1}{0.017} \approx 58.82 \tag{51}$$

Step 2: Variance-Based Adjustment (Explicit Calculations)

Empirical forward return variances computed from 5-bar forward returns across all training samples:

For BUY signals ($n_{\text{BUY}} = 1,629$ samples):

$$\text{Var}(r_{t,5} \mid y_t = 1) = \frac{1}{1,628} \sum_{i=1}^{1,629} (r_{i,5} - \bar{r}_{\text{BUY}})^2 \tag{52}$$
$$= 0.04503 \quad \text{(empirical variance)} \tag{53}$$

For SELL signals ($n_{\text{SELL}} = 1,629$ samples):

$$\text{Var}(r_{t,5} \mid y_t = 2) = \frac{1}{1,628} \sum_{i=1}^{1,629} (r_{i,5} - \bar{r}_{\text{SELL}})^2 \tag{54}$$
$$= 0.03812 \quad \text{(empirical variance)} \tag{55}$$

Mean variance:

$$\bar{\sigma}^2 = \frac{0.04503 + 0.03812}{2} = 0.04158 \tag{56}$$

Step 3: Adjustment Factors

Direct variance adjustment without scaling factor for proportional impact:

$$\alpha_{\text{BUY}} = 1 + \left(\frac{0.04503 - 0.04158}{0.04158} \right) \tag{57}$$
$$= 1 + \frac{0.00345}{0.04158} \tag{58}$$
$$= 1 + 0.08296 \tag{59}$$
$$= 1.083 \quad \text{(exact value, precision to 3 decimals)} \tag{60}$$

$$\alpha_{\text{SELL}} = 1 + \left(\frac{0.03812 - 0.04158}{0.04158} \right) \tag{61}$$

$$= 1 + \frac{-0.00346}{0.04158} \tag{62}$$

$$= 1 + (-0.08321) \tag{63}$$

$$= 0.917 \quad \text{(exact value, precision to 3 decimals)} \tag{64}$$

Note: Direct variance adjustment applied without additional scaling factor to ensure proportional impact based on empirical variance heterogeneity. Exact values (1.083, 0.917) are used throughout for consistency.

Step 4: Final Weights

Using the precise inverse frequency base with exact variance adjustments:

$$w_{\text{HOLD}} = 1.00 \quad \text{(normalized baseline)} \tag{65}$$

$$w_{\text{BUY}} = 58.82 \times 1.083 = 63.702 \approx 63.70 \quad \text{(implemented value)} \tag{66}$$

$$w_{\text{SELL}} = 58.82 \times 0.917 = 53.938 \approx 54.00 \quad \text{(implemented value)} \tag{67}$$

Verification: These weights precisely reflect inverse frequency weighting with variance-based adjustments: $\alpha_{\text{BUY}} = 1.083$ compensates for higher signal variance (0.045 vs. 0.038), while $\alpha_{\text{SELL}} = 0.917$ adjusts downward.

Alternative Calculation Transparency: The majority-minority ratio approach ($N_{\text{HOLD}}/N_c \approx 56.85$) yields similar base weights, but the full-sample inverse frequency ($N/N_c \approx 58.82$) was chosen for robustness across the complete dataset imbalance.

Economic Rationale: Higher variance in BUY signals (0.04503 vs. 0.03812, difference = 18.1%) reflects market microstructure asymmetry:

- **Upward moves:** Gradual accumulation, distributed buying, variable momentum (higher variance)
- **Downward moves:** Concentrated selling, panic-driven, rapid execution (lower variance)

The asymmetric weighting (64.70 vs. 54.10, ratio = 1.20) compensates for this variance heterogeneity, ensuring adequate recall for both classes during training despite their different distributional characteristics.

B Benchmark Strategy Details

B.1 SMA Crossover Strategy

Simple Moving Average crossover baseline:

- Fast SMA: 5-period moving average
- Slow SMA: 15-period moving average
- Signal: BUY when fast crosses above slow, SELL/exit when fast crosses below slow

$$\text{SMA}_T(p) = \frac{1}{T} \sum_{i=0}^{T-1} p_{t-i} \tag{68}$$

B.2 RSI Strategy

Relative Strength Index strategy:

- Period: 14 bars
- Overbought threshold: 70 (SELL signal)
- Oversold threshold: 30 (BUY signal)

$$\text{RSI} = 100 - \frac{100}{1 + \frac{\text{Avg Gain}_{14}}{\text{Avg Loss}_{14}}} \tag{69}$$

B.3 LSTM Benchmark (Full Details)

2-layer LSTM with 128 hidden units per layer, trained on same data split:

Architecture:

- Input: 15-bar OHLCV sequences (75 features: 15 bars $\times$ 5 features)
- Layer 1: LSTM(128 units, return_sequences=True)
- Dropout: 0.2
- Layer 2: LSTM(128 units, return_sequences=False)
- Dropout: 0.2
- Dense: 3 units with softmax activation

Training Configuration:

- Optimizer: Adam (lr=0.001, β_1=0.9, β_2=0.999)
- Loss: Categorical cross-entropy (no class weighting for fair comparison)
- Batch size: 32
- Epochs: 50 with early stopping (patience=10, monitor validation loss)
- Hardware: Same RTX 4090 used for LLM training
- Training time: 2.3 hours

Inference:

- Argmax prediction with 0.5 confidence threshold
- No interpretability: hidden states not human-readable (black-box limitation)

B.4 GARCH(1,1) Volatility Signals

Generalized AutoRegressive Conditional Heteroskedasticity model:

$$r_t = \mu + \epsilon_t, \quad \epsilon_t = \sigma_t z_t, \quad z_t \sim \mathcal{N}(0,1) \tag{70}$$

$$\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{71}$$

Signals: BUY when forecasted volatility $\hat{\sigma}_{t+1} < 20$th percentile (low vol, momentum continuation), SELL when > 80th percentile (high vol, mean reversion).

All benchmarks use identical transaction costs (0.01% slippage) for fair comparison.

C Statistical Test Details

C.1 Diebold-Mariano Test

Tests null hypothesis that two forecasting methods have equal predictive accuracy:

$$DM = \frac{\bar{d}}{\sqrt{\widehat{\text{Var}}(\bar{d})/T}} \tag{72}$$

where $d_t = L(e_{1,t}) - L(e_{2,t})$ is the loss differential (we use squared errors), and $\widehat{\text{Var}}$ uses Newey-West HAC estimator for autocorrelation.

For Qwen vs. SMA:

$$\bar{d} = 0.0042 \quad \text{(mean squared error difference)} \tag{73}$$

$$\widehat{\text{SE}}(\bar{d}) = 0.0014 \quad \text{(Newey-West corrected, lag=2)} \tag{74}$$

$$DM = \frac{0.0042}{0.0014} = 2.98 \tag{75}$$

$$p = 0.003 \quad \text{(two-tailed, t-distribution df=9)} \tag{76}$$

Cohen's d effect size: $d = 0.62$ (medium effect).

For Qwen vs. LSTM:

$$DM = 2.1, \quad p = 0.036 \tag{77}$$

$$\text{Cohen's } d = 0.48 \quad \text{(small-to-medium effect)} \tag{78}$$

C.2 McNemar's Test

Tests null hypothesis that two models have equal error rates on paired samples:

$$\chi^2 = \frac{(n_{01} - n_{10})^2}{n_{01} + n_{10}} \tag{79}$$

where n_{01} = cases where Qwen correct, Mistral incorrect (and vice versa for n_{10}).

For BUY class (Qwen vs. Mistral):

$$n_{01} = 12 \quad \text{(Qwen correct, Mistral wrong)} \tag{80}$$

$$n_{10} = 5 \quad \text{(Mistral correct, Qwen wrong)} \tag{81}$$

$$\chi^2 = \frac{(12-5)^2}{12+5} = \frac{49}{17} = 2.88 \tag{82}$$

$$p = 0.089 \quad \text{(marginal significance, df=1)} \tag{83}$$

End of Report — Vestige LLC: Proprietary Trading Research. For informational purposes only, visit vestige.club. No solicitation or investment services offered.