QuantAgent: Price-Driven Multi-Agent LLMs for High-Frequency Trading

Fei Xiong[1,2,†], **Xiang Zhang**[3,†], **Aosong Feng**[4], **Siqi Sun**[5], **Chenyu You**[1]

[1]Stony Brook University, [2]Carnegie Mellon University, [3]University of British Columbia, [4]Yale University, [5]Fudan University

[†]Equal contribution

Recent advances in Large Language Models (LLMs) have shown remarkable capabilities in financial reasoning and market understanding. Multi-agent LLM frameworks such as TradingAgent and FINMEM augment these models to long-horizon investment tasks by leveraging fundamental and sentiment-based inputs for strategic decision-making. However, these approaches are ill-suited for the high-speed, precision-critical demands of *High-Frequency Trading* (HFT). HFT typically requires rapid, risk-aware decisions driven by structured, short-horizon signals, such as technical indicators, chart patterns, and trend features. These signals stand in sharp contrast to the long-horizon, text-driven reasoning that characterizes most existing LLM-based systems in finance. To bridge this gap, we introduce **QuantAgent**, the first multi-agent LLM framework explicitly designed for high-frequency algorithmic trading. The system decomposes trading into four specialized agents, *Indicator*, *Pattern*, *Trend*, and *Risk*, each equipped with domain-specific tools and structured reasoning capabilities to capture distinct aspects of market dynamics over short temporal windows. Extensive experiments across nine financial instruments, including Bitcoin and Nasdaq futures, demonstrate that QuantAgent consistently outperforms baseline methods, achieving higher predictive accuracy at both 1-hour and 4-hour trading intervals across multiple evaluation metrics. Our findings suggest that coupling structured trading signals with LLM-based reasoning provides a viable path for traceable, real-time decision systems in high-frequency financial markets.

Github: https://github.com/Y-Research-SBU/QuantAgent

Website: https://Y-Research-SBU.github.io/QuantAgent/

Corresponding Authors: chenyu.you@stonybrook.edu, siqisun@fudan.edu.cn

1 Introduction

In quantitative finance, technical analysis treats historical price action as the most immediate and information-dense reflection of market conditions (Pring, 1991). The central premise is that market dynamics, including fundamentals, macro events, institutional flows, and collective sentiment, are ultimately embedded in price movements (Murphy, 1999). Each bar, defined by its open, high, low, and close (OHLC), provides a compact yet universal representation of short-horizon market behavior. This structure enables systematic detection of recurring setups such as trends, reversals, breakouts, and momentum shifts across asset classes ranging from equities and commodities to digital assets (Moskowitz et al., 2012). Under the efficient market hypothesis (Fama, 1970), prices adjust rapidly to public information, making patterns in OHLC bars a natural substrate for short-term prediction without reliance on lagging textual inputs.

Large Language Models (LLMs) have recently demonstrated impressive capabilities in multi-step reasoning, tool use, and interpretable decision-making (OpenAI et al., 2024). These capabilities are directly relevant to quantitative trading (Yang et al., 2023), which heavily depends on integrating heterogeneous signals, applying systematic trading rules, and controlling execution risks. However, most existing LLM-driven financial frameworks operate primarily on textual inputs, such as news articles, social media streams, or earnings reports (Nguyen et al., 2015; Xiao et al., 2025; Zakir et al., 2025; Zhang et al., 2023a). This reliance introduces two major limitations: (i) textual signals typically lag price discovery and are incorporated into markets only after the fact (Chordia et al., 2013), and (ii) such data is noisy, unstructured, and difficult to validate (Liu et al., 2022a). Since short-horizon market dynamics are already encoded in OHLC bars, a more direct approach

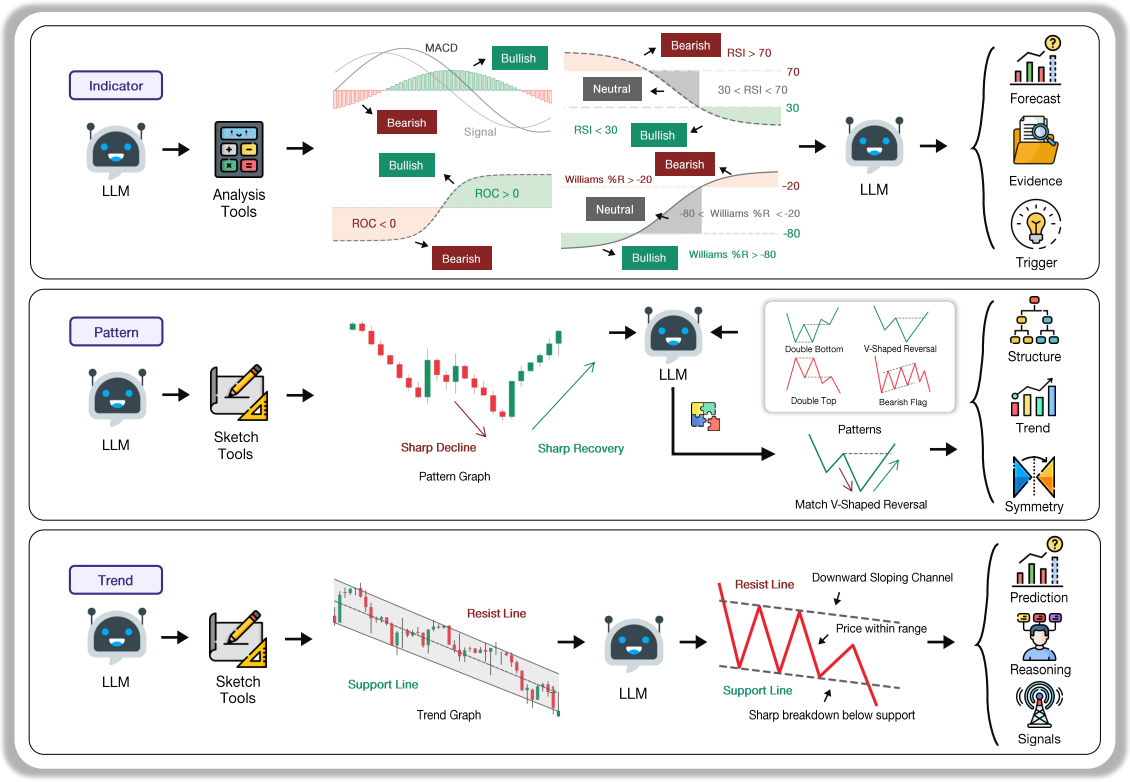

Figure 1 Workflows of IndicatorAgent, PatternAgent, and TrendAgent. IndicatorAgent interprets signals from MACD, RSI, ROC, and Williams %R; PatternAgent detects formations such as double bottoms; TrendAgent extracts directional flow via support and resistance channels.

is to align LLM reasoning with structured price-based signals. To the best of our knowledge, no prior work has developed an LLM-based framework for high-frequency trading (HFT) that operates directly on OHLC data.

In this paper, we propose **QuantAgent** (Figure 1), the first multi-agent LLM framework tailored to high-frequency algorithmic trading. Specifically, QuantAgent decomposes the trading process into four specialized agents – IndicatorAgent, PatternAgent, TrendAgent, and RiskAgent – each designed to capture a complementary dimension of technical analysis. IndicatorAgent condenses raw OHLC bars into robust technical indicators, providing a noise-resistant summary of recent market behavior. PatternAgent chart formations such as peaks, troughs, and consolidations, leveraging the multimodal reasoning abilities of LLMs (Nison, 2001). TrendAgent identifies directional bias from short-horizon price dynamics, while RiskAgent integrates all signals into a coherent risk–reward profile. Final trade decisions emerge from the interaction of these agents, yielding traceable, language-native rationales that can be inspected alongside execution (Schick et al., 2023).

We evaluate QuantAgent on a multi-asset benchmark spanning commodities, equities, cryptocurrencies, and volatility indices. At 1-hour and 4-hour bar resolutions, QuantAgent consistently outperforms baselines across both directional accuracy and return-based metrics, with particularly pronounced gains in equity markets. Rolling-window validation further demonstrates robust generalization, achieving up to 80% directional accuracy in forecasting short-term price movements. Besides its strong empirical performance, QuantAgent provides natural-language rationales for trading decisions, enabling a degree of traceability and interpretability often missing in traditional algorithmic strategies.

2 Related Works

Agent-Based LLMs for Financial Decision-Making. The design of QuantAgent builds on recent work that organizes LLMs into multi-agent systems for financial decision-making. FINCON (Yu et al., 2024) introduces a manager–analyst hierarchy trained via verbal reinforcement learning, while TradingAgents (Xiao et al.,

2025) models institutional workflows through agent communication, prioritizing interpretability over the low-latency, price-driven logic required in high-frequency trading (Tumarkin and Whitelaw, 2001). As a line of work, these systems demonstrate the potential of LLM-based agents in finance, but their heavy reliance on textual inputs leaves them ill-suited for the structured, low-latency signals required in HFT scenarios. More recently, RD-Agent(Q) (Li et al., 2025) takes a significant step forward by shifting to structured, data-centric signals and automating factor–model co-optimization.However, RD-Agent(Q) remains constrained to daily-resolution strategies and slower research-feedback cycles, making it less suitable for real-time decision-making in high-frequency contexts.

Quantitative Trading Based on Indicators and Patterns. Prior to LLM-based agents, quantitative trading systems are predominantly built on technical indicators such as trends, volatility, and momentum for intraday decision-making. Early studies show that nonlinear price patterns can exhibit predictive power (Lo et al., 2000), but subsequent work highlight challenges including overfitting and researcher bias (Chen and Chen, 2016). Momentum strategies (Jegadeesh and Titman, 1993; Moskowitz et al., 2012) are widely adopted to capture trend persistence, while heuristics such as Elliott wave theory (Prechter, 2005) and curated pattern libraries attempt to model higher-order market structures. Although these indicator- and pattern-based approaches are interpretable and computationally efficient, they often struggle in volatile or noisy environments, undermining their effectiveness in high-frequency settings. These limitations motivate us to design a framework that fuses structured price signals with LLM-based reasoning, enabling more adaptive and interpretable trading systems.

3 QuantAgent

To bridge the gap between traditional high-frequency quantitative trading and recent advances in multi-agent LLM systems, we introduce QuantAgent, a collaborative framework for low-latency market decision-making. QuantAgent integrates classical technical analysis with prompt-structured LLM reasoning, enabling modular and interpretable financial intelligence. Built on LangGraph (LangChain, 2025), the system simulates the workflow of institutional trading desks, where specialized agents execute distinct analytical roles to support rapid and coordinated decision-making.

In contrast to prior LLM-based frameworks that incorporate external sources such as news or social media sentiment, QuantAgent operates solely on price-derived market signals. This design choice reflects the efficient market hypothesis, which posits that asset prices incorporate available information by aggregating the actions and beliefs of market participants over time (Murphy, 1999). By grounding analysis exclusively in OHLC data and technical indicators, QuantAgent avoids the latency, noise, and unpredictability of textual sources, while remaining fast, interpretable, and directly aligned with the demands of high-frequency trading.

The system decomposes trading into four specialized agent, IndicatorAgent, PatternAgent, TrendAgent, and RiskAgent, that communicate through structured prompts. Each agent captures a complementary perspective on short-horizon market dynamics: numerical indicators, geometric patterns, directional momentum, and integrated decision-making. In the following subsections, we describe the design of each agent in detail and formalize the technical components underlying our framework.

Algorithm 1: Slope-aware trend detection over a candlestick sequence $P_{0:T-1}$

Input: $P_{0:T-1}$, N, τ

for $t = N-1$ **to** $T-1$ **do**
  Fit OLS on highs/lows to get m_r, m_s;
  $\kappa_t \leftarrow (m_r + m_s)/2$;
  if $\kappa_t > \tau$ **then** Trend $\leftarrow$ Uptrend;
  else if $\kappa_t < -\tau$ **then** Trend $\leftarrow$ Downtrend;
  else Trend $\leftarrow$ Sideways;
end for
Render chart $\mathcal{K}_t(P_t, \kappa_t, \text{Trend})$;