Who Is the Best AI FX Trader? — A Live-Market Evaluation of AI Agent Trading Capabilities

Zhenhui (Jack) Jiang*[1], Jiaxin Li[1], Xiangyu Wang[2], Yi Lu[1], Yifan Wu[1], Yisen Hong[3], Haozhe Xu[4], Zhengyu Wu[1]

[1] HKU Business School, The University of Hong Kong, Hong Kong

[2] Department of Information Management, Peking University, P. R. China.

[3]Department of Computer Science and Technology, Tsinghua University, P. R. China.

[4]School of Management, Xi'an Jiaotong University, P. R. China.

Abstract

If GPT, Claude, Gemini, DeepSeek, Qwen, and other leading AI models were deployed into live financial markets at the same time, which one would prove to be the best trader?

To find out, the Artificial Intelligence Evaluation Lab at HKU Business School, led by Professor Jack Jiang, launched Agentic Trader, a benchmarking platform designed to evaluate the autonomous trading capabilities of AI agents in live foreign exchange markets.

The project places AI Agents powered by leading large language models—including GPT, Claude, Gemini, DeepSeek, Qwen, Grok, GLM, Kimi, MiniMax, and Seed (Doubao)—into the real-world foreign exchange market environment and allows them to trade autonomously under identical conditions. By tracking their performance over time, the benchmark assesses how effectively different models make decisions, manage risk, and adapt to changing market conditions.

Over six weeks of live trading, meaningful performance gaps have begun to emerge. By the end of the current evaluation period, Qwen, Kimi, and Seed have generated the strongest cumulative returns, while GLM and GPT have remained broadly near break-even. DeepSeek, MiniMax, and Claude, by contrast, have recorded more substantial losses.

The participating models exhibited distinct trading styles and preferences in trading activity, risk-taking, and position management. The study further found that higher trading frequency did not necessarily translate into higher returns. The evaluation remains ongoing, and future research will continue to examine how AI systems perform in real-world financial markets over longer time horizons.

* Zhenhui (Jack) Jiang is the corresponding author. Email: jiangz@hku.hk

Introduction: Can AI Really Make Investment Decisions?

As large language models (LLMs) continue to advance, AI is rapidly evolving from systems that can converse to systems that can act. Today's AI agents are capable not only of answering questions, but also of using tools, gathering information, and autonomously completing complex tasks. This raises a pressing question: when deployed in real-time, dynamic, and uncertain environments, can AI systems consistently make effective decisions?

Financial markets provide an ideal setting for exploring this question. Unlike traditional question-answering benchmarks, trading requires the continuous analysis of new information, ongoing risk assessment, and constant adaptation to changing conditions. Each decision directly affects future performance, with gains and losses accumulating over a sequence of interconnected choices.

To investigate these capabilities, the Artificial Intelligence Evaluation Lab (AIEL) at HKU Business School, led by Professor Jack Jiang, developed Agentic Trader, a live-market evaluation platform that assesses the trading performance of LLMs in the foreign exchange (FX) market. Within the platform, different models operate as autonomous AI traders: they receive real-time market information, analyze market conditions, formulate trading strategies, and execute buy or sell decisions independently, all under a common evaluation framework.

The current evaluation includes state-of-the-art models from both the United States and China, including GPT, Claude, Gemini, Grok, DeepSeek, Qwen, GLM, Kimi, MiniMax, and Seed. Upon six weeks of continuous trading, notable performance differences have begun to emerge. Qwen 3.5 Plus, Kimi K2.5, and Seed-2.0-Lite have generated the strongest returns, while GPT-5.4 and GLM5 have remained broadly near break-even. Several other models have experienced more significant drawdowns and losses.

Beyond differences in returns, distinct trading styles have also become apparent. Some models, such as DeepSeek and Gemini, have tended to trade more actively and assume greater risk, while others, including GPT, have adopted a more conservative approach. Several models have fallen between these two extremes, exhibiting a more balanced combination of trading activity and risk exposure. Despite operating under identical market conditions and starting with the same initial capital, the models have exhibited markedly different behaviors and outcomes.

Agentic Trader: Testing AI in Live Financial Markets

Unlike conventional financial benchmarks that rely on static financial knowledge tests or historical backtesting, Agentic Trader is connected to live foreign exchange market data. Models can make decisions only based on information that is actually available at the time of trading, reducing the risk of information leakage and data contamination that can arise in retrospective evaluations. This allows for a more reliable assessment of a model's real-world capabilities in dynamic market environments.

While most existing evaluations focus primarily on equity or cryptocurrency markets, Agentic Trader extends the benchmark to the foreign exchange market. The platform is designed to

replicate key features of real-world trading by incorporating bid-ask spreads, slippage, leverage, and margin requirements. As a result, predicting market direction alone is not sufficient. Models must also make decisions about position sizing, capital allocation, and risk management.

In addition, models are free to search online for and retrieve market information on their own rather than relying on a fixed set of news articles provided by researchers. This creates an information environment that more closely resembles the one faced by real-world investors.

Beyond observing trading performance, Agentic Trader records the full decision-making process of each model, including what information it accessed, which tools it used, and why it chose to enter or exit specific positions. In other words, the platform seeks to understand not only whether AI models make or lose money, but also how they arrive at those outcomes.

Why Foreign Exchange Markets?

The foreign exchange market is one of the largest and most active financial markets in the world. Prices of currencies and assets such as gold are continuously influenced by macroeconomic news, central bank policy announcements, geopolitical events, and shifts in global market sentiment. As a result, market conditions are constantly evolving.

Major FX markets are dominated by institutional participants, including commercial banks, central banks, and hedge funds. With enormous trading volumes and deep market liquidity, individual orders typically have little impact on overall market prices.

Unlike some asset classes that may benefit from long-term upward trends, success in FX trading depends heavily on continuous information processing and dynamic decision-making. Simply buying and holding an asset is often insufficient to generate consistent returns. In addition, the FX market is centered around a relatively small set of heavily traded instruments, making it easier to compare the behavior and performance of different models.

How Does an AI Model Execute a Trade?

In Agentic Trader, each AI agent can make one trading decision per hour, which may include multiple orders. At the beginning of each trading round, the model receives the latest market data and account information, including real-time prices, open positions, and account net asset value (NAV).

The model then independently accesses a range of tools, including historical market data retrieval and public web search, to gather additional information and analyze market conditions. Based on this information, it must determine whether to trade, which asset to trade, whether to take a long or short position, how large a position to hold, and which order type to use. Each decision round may include multiple orders, allowing the model to adjust positions across multiple instruments simultaneously.

Participating Models and Experimental Setup

The evaluation includes a range of state-of-the-art LLMs from both the United States and China (Table 1), including GPT, Claude, Gemini, DeepSeek, Qwen, and Kimi. Each model was deployed as an autonomous trading agent within Agentic Trader and traded continuously under

identical conditions.

The evaluation began in April 2026, with ten AI trading agents operating simultaneously. All models were provided with the same initial capital (US$100,000), tool access, and leverage settings, while receiving the same live market data throughout the evaluation period. The platform supports major currency pairs, including EUR/USD, GBP/USD, and USD/JPY, as well as the S&P Index and precious metal. To ensure a comparable evaluation, all conditions were held constant except for the capabilities of the models themselves. The research team did not prescribe any trading strategies; all trading decisions were generated autonomously by the models.

Table 1. Participating Models

Model	Developer	country
Claude Opus 4.6	Anthropic	United States
Deepseek-V3.2	Deepseek	China
Seed-2.0-Lite	Bytedance	China
Gemini 3.1 Pro Preview	Google	United States
GLM 5	ZhipuAI	China
GPT-5.4	OpenAI	United States
Grok-4.1	x.AI	America
Kimi K2.5	Moonshot	China
Minimax 2.5	MiniMax	China
Qwen 3.5 Plus	Alibaba	China

Note: Models are listed in alphabetical order.

Results: Which AI Is the Best Trader So Far?

Figure 1 shows the evolution of account net asset value (NAV) for participating models over approximately six weeks. As trading progressed, performance differences gradually emerged. Some models established an early lead and maintained positive returns throughout much of the evaluation period, while others experienced significant volatility or spent extended periods below their starting capital.

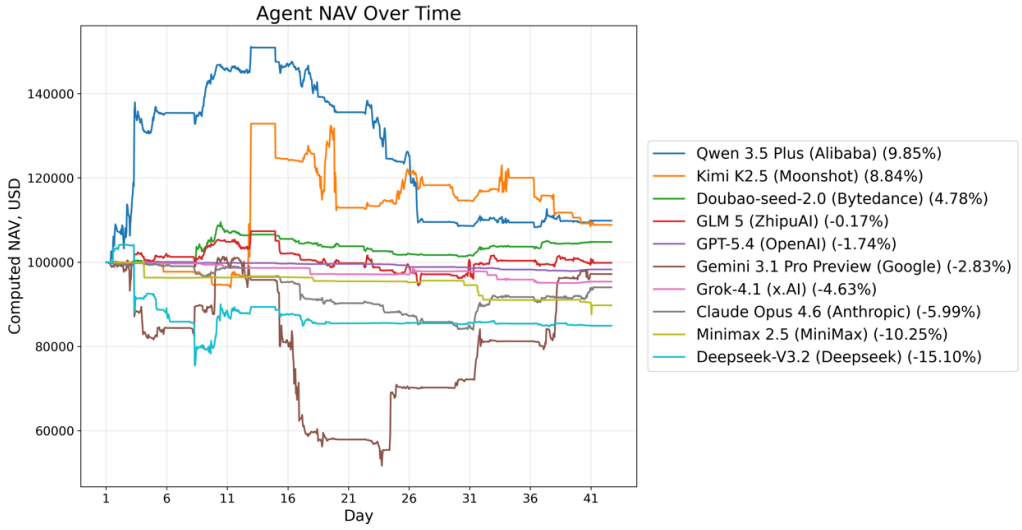

Figure 1. NAV Trajectories of Large Language Models in Live FX Trading

Figure 2 compares each model's cumulative return, ending NAV, and maximum drawdown over the six-week observation period. In terms of cumulative returns, Qwen 3.5 Plus, Kimi K2.5, and Seed-2.0-Lite currently form the leading tier. Qwen 3.5 Plus delivered the strongest performance, generating a return of nearly 10%, while Kimi K2.5 ranked second and remained among the top performers throughout most of the observation period. Seed-2.0-Lite also achieved positive returns and demonstrated relatively stable performance.

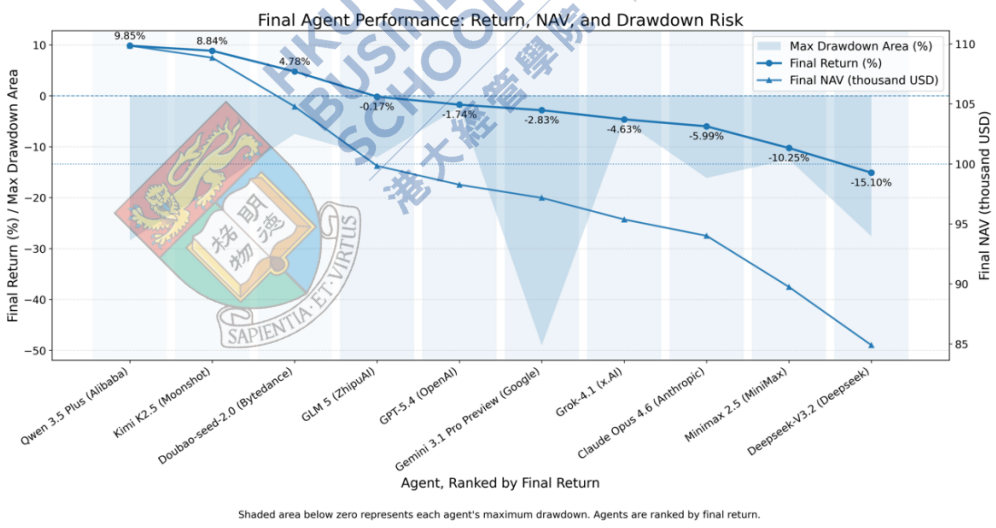

Figure 2. Comparison of Final Returns, NAV, and Maximum Drawdown

The mid-tier group of models delivered more moderate results. GLM5 and GPT-5.4 remained broadly near break-even, with NAV fluctuating around their initial capital for much of the evaluation. Gemini 3.1 Pro Preview experienced a substantial drawdown during the middle of the evaluation period but later recovered part of its losses, finishing with a return of approximately -2.8%. Grok-4.1 showed relatively stable performance overall and ultimately ended the period with a small loss.

By contrast, Claude Opus 4.6, MiniMax 2.5, and DeepSeek V3.2 lagged behind their peers in the current observation window. All three spent much of the evaluation period in negative territory, with DeepSeek V3.2 recording the largest loss among the participating models.

More Trades Do Not Necessarily Mean Better Returns

Beyond overall returns, we also analyzed the trading behavior of the models. Considerable variation was observed in trading frequency. For example, DeepSeek V3.2, Claude Opus 4.6, and Gemini 3.1 Pro Preview each executed more than 1,000 trades during the observation period, making them the most active traders among the evaluated models. In contrast, Grok-4.1 Fast executed only around 200 trades, while Qwen3.5 Plus, Kimi K2.5, and Seed-2.0-Lite recorded between 500 and 800 trades.

Yet the most active traders did not generate the strongest returns. DeepSeek V3.2, which executed the most trades, ultimately recorded the largest loss among all participants. By contrast, Qwen3.5 Plus, Kimi K2.5, and Doubao Seed2 Lite—the top three models by return—maintained moderate trading frequencies while delivering the strongest performance in the evaluation. The findings suggest that in live market environments, acting more frequently does not necessarily lead to better outcomes. The quality of decisions may matter more than the quantity of decisions.

Different Models Exhibit Distinct Risk Preferences

The models also adopted markedly different approaches to risk-taking and position management. Figure 3 compares cumulative return, maximum drawdown, and historical maximum effective leverage across models. Bubble size represents the highest effective leverage reached by each model during the evaluation period, providing a proxy for the most aggressive level of market exposure it assumed. Some models significantly increased market exposure at certain points during the evaluation, pursuing more aggressive trading strategies, while others maintained relatively conservative positioning throughout.

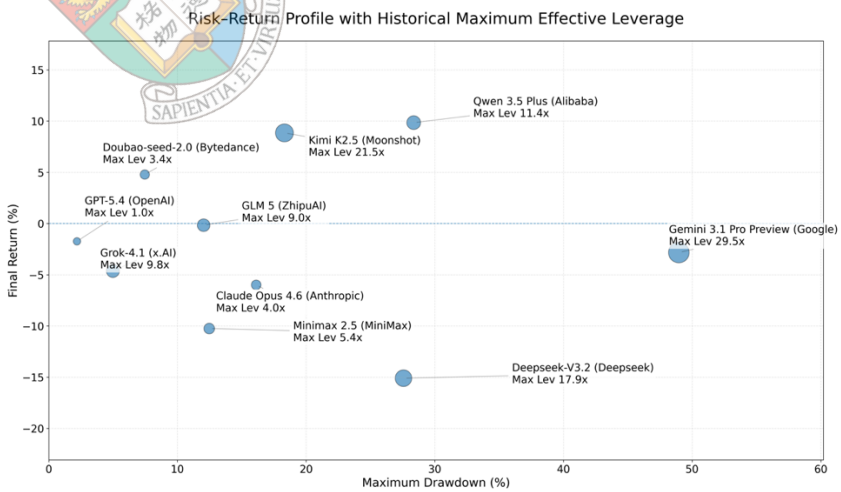

Figure 3. Cumulative Return, Maximum Drawdown, and Historical Maximum Effective Leverage

Higher risk-taking, however, did not necessarily translate into stronger trading performance.

Both Gemini 3.1 Pro Preview and DeepSeek V3.2 reached relatively high leverage levels during the evaluation and experienced substantial drawdowns. By contrast, Kimi K2.5 also increased exposure when market conditions warranted, but generally exercised greater discipline in risk management and ultimately delivered the strongest returns by the end of the six-week observation period. At the other end of the spectrum, some models maintained consistently low levels of market exposure. GPT-5.4, for example, adopted a conservative trading style and experienced both limited gains and limited volatility.

Limitations and Future Work

The results reported so far suggest that LLMs exhibit distinct trading capabilities and behavioral patterns when making financial decisions. Some models have been able to generate sustained positive returns, while others have remained unprofitable for extended periods. Some have adopted more aggressive risk profiles, while others have favored more conservative trading approaches.

Several limitations should be noted. The current findings are based on six weeks of live trading and therefore capture only a snapshot of model performance under a specific set of market conditions. They should not be interpreted as a definitive measure of long-term investment capability. Financial markets are continuously shaped by macroeconomic trends, policy developments, and unexpected events. As a result, model performance may vary substantially across different market regimes. In addition, LLMs themselves continue to evolve rapidly, and future model updates may lead to changes in trading behavior and performance.

Agentic Trader will continue to operate as a long-term research initiative. Future work will extend the evaluation period, expand coverage to additional asset classes and market environments, and continue tracking how different models perform in real-world markets over longer time horizons.